Parameter identifiability of marine ecosystem models

Outline (lecture, 45 min)

1) Dynamical, process-based modelling to derive mass flux estimates

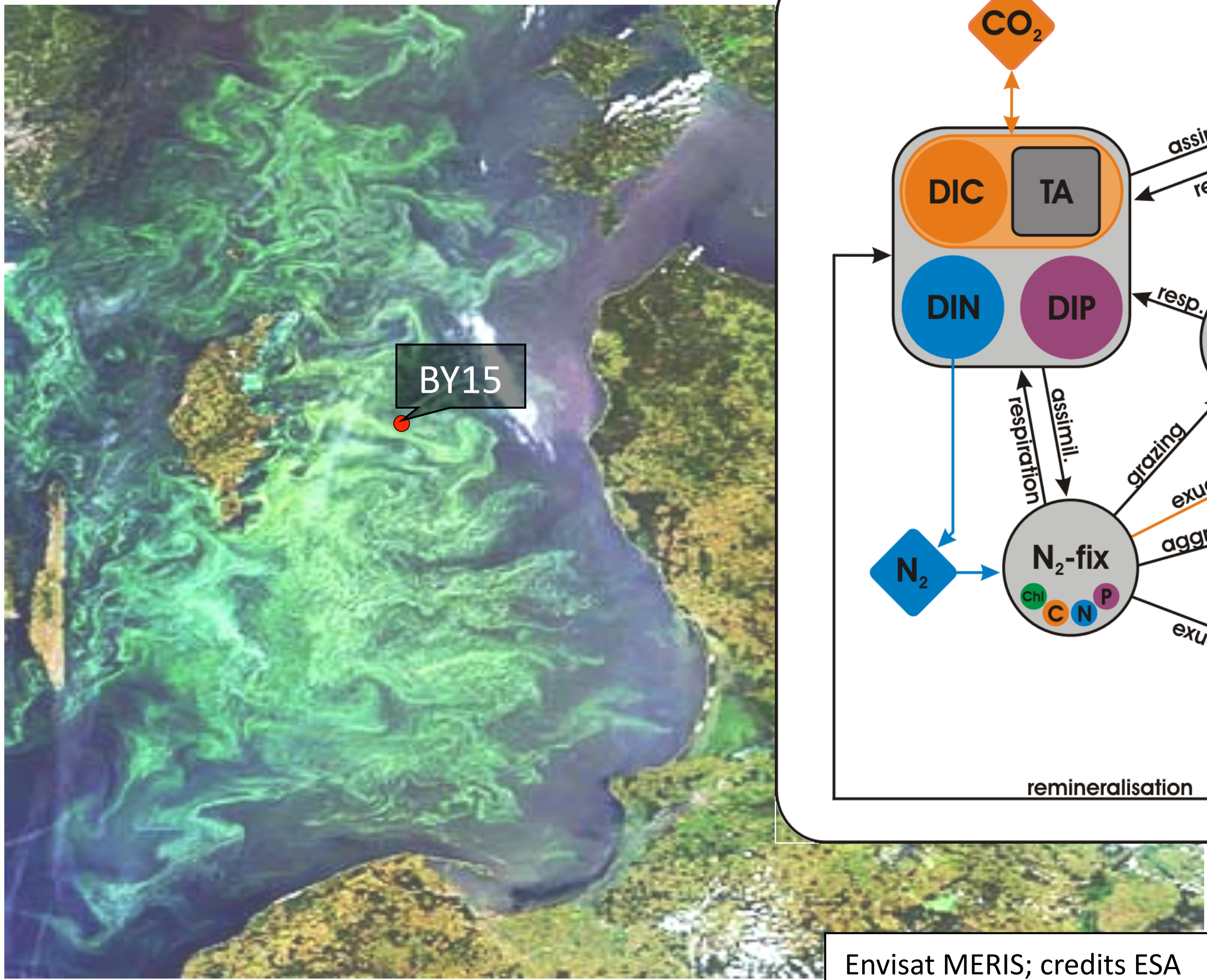
2) Parameter estimation (basics *again* recalled)

3) What is meant by *parameter identifiability*?

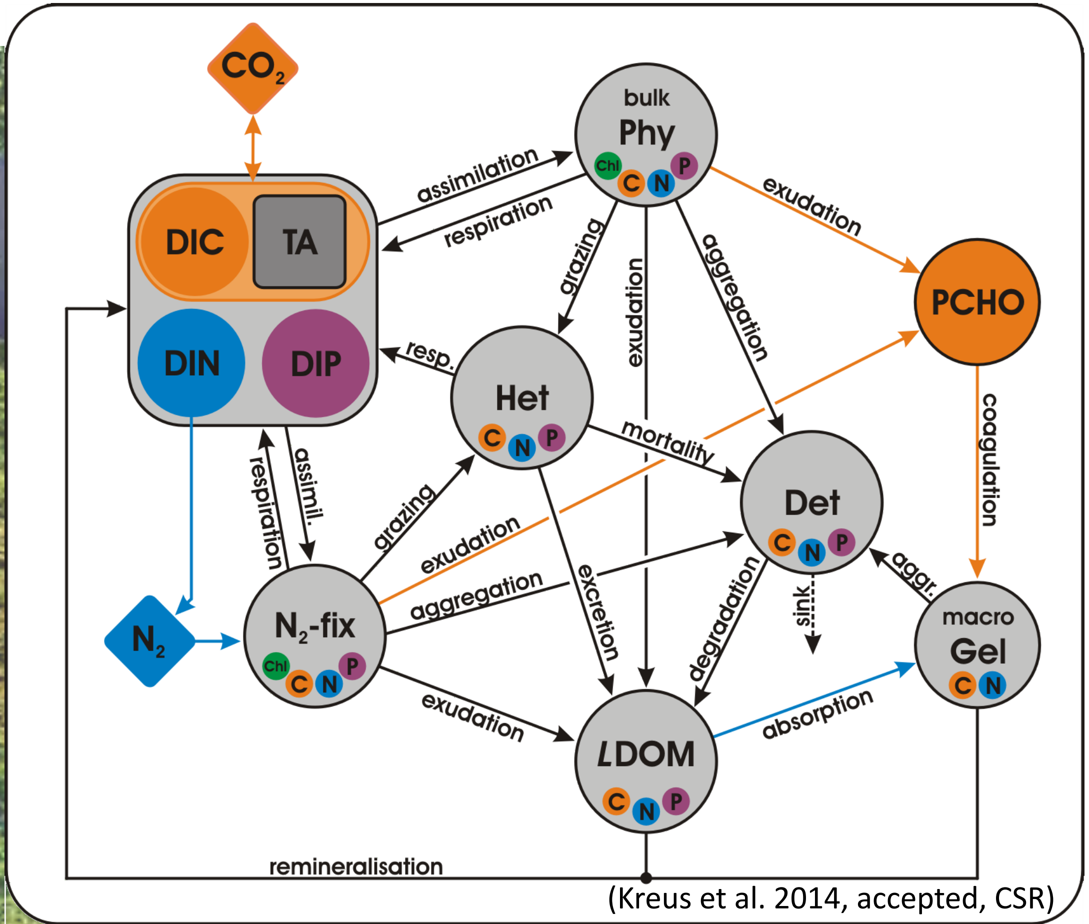
4) Resampling strategy to specify confidence-, credibility regions in parameter space

1) Dynamical, process-based modelling to derive mass flux estimates

Example: 1D-(vertically resolved) plankton ecosystem model at local site in central Baltic Sea



Envisat MERIS; credits ESA



(Kreus et al. 2014, accepted, CSR)

1) Dynamical, process-based modelling to derive mass flux estimates

Model state variables (x) at times (t_i) with dynamics that also depends on environmental variables (I):

$$\mathbf{x}(t_{i+1}) = M_i \left[\mathbf{x}(t_i), \theta, I\right]$$

Observation vector (observable, y^o) at times (t_i), with observation operator (H_i) and system noise (variability, ε_i):

$$\mathbf{y}^o(t_i) = H_i \left[\mathbf{x}(t_i)\right] + \epsilon_i$$

1) Dynamical, process-based modelling to derive mass flux estimates

Model state variables (x) at times (t_i):

$$\mathrm{x}(t_{i+1}) = M_i\left[\mathrm{x}(t_i), \theta, I\right]$$

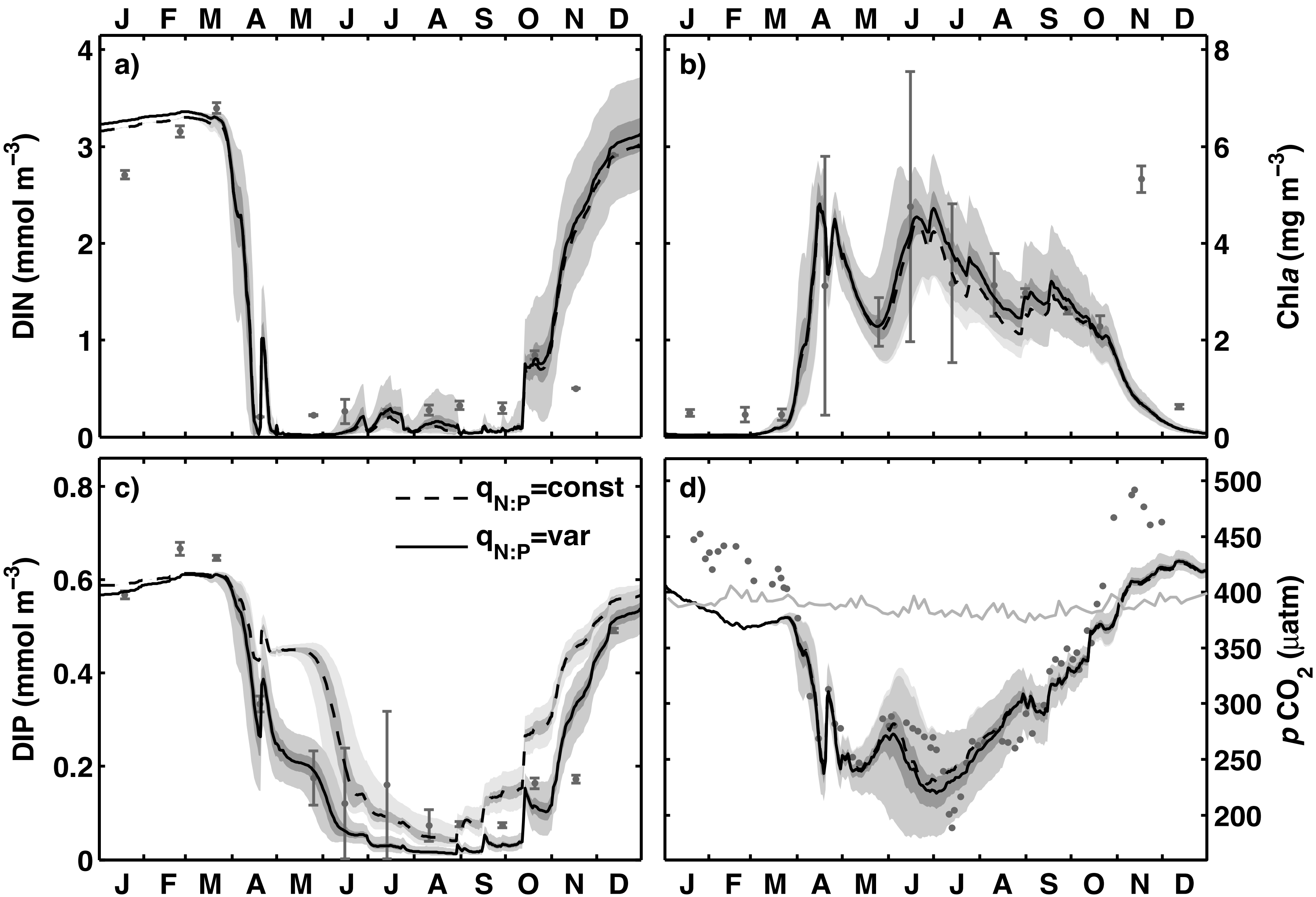
Observation vector (observable, y^o) at times (t_i), with observation operator (H_i) and system noise (variability, ε_i):

$$\mathrm{y}^o(t_i) = H_i\left[\mathrm{x}(t_i)\right] + \epsilon_i$$

with (for example)

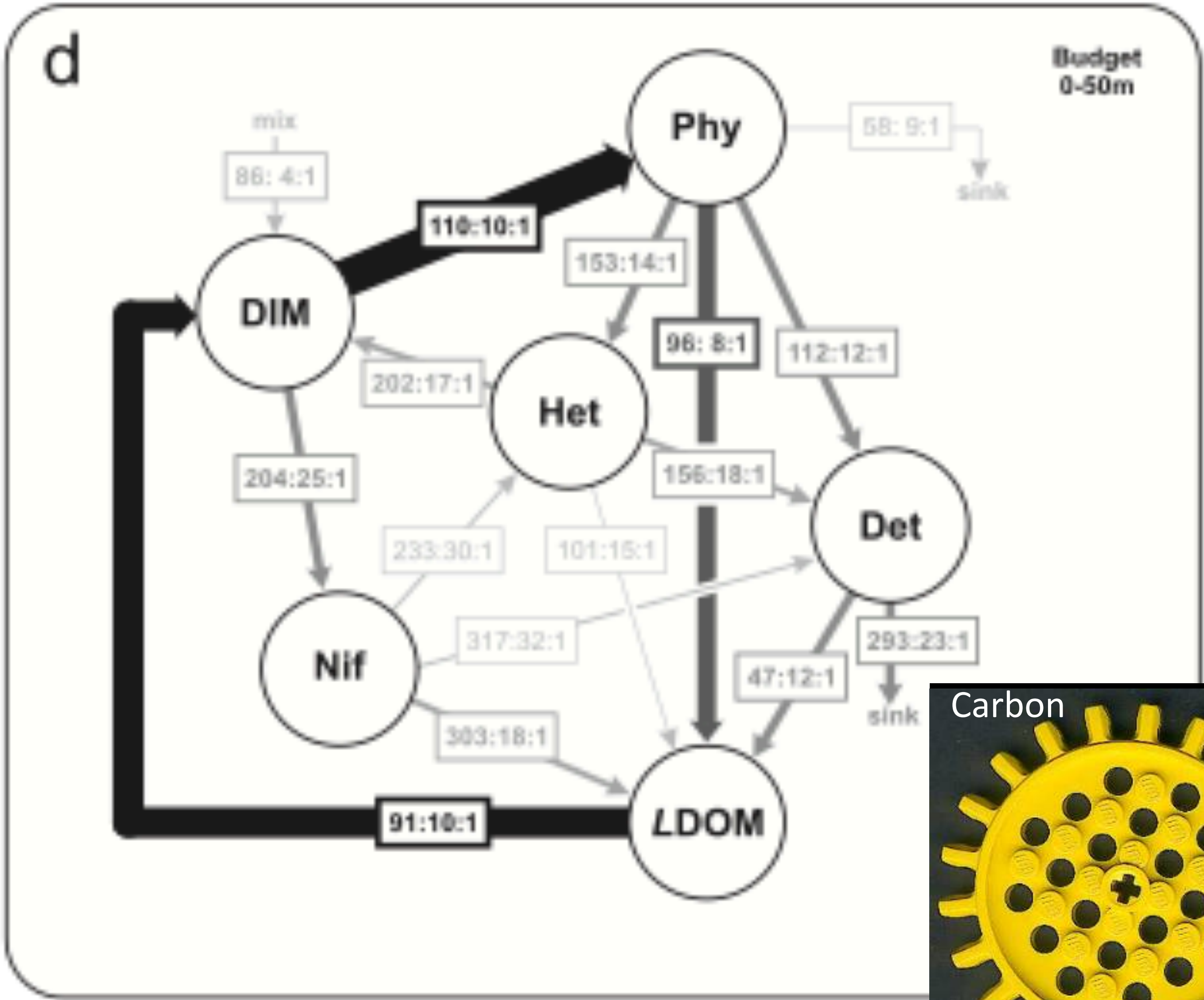
$$H_i = \left(\begin{array}{l} (p\mathrm{CO}_2)_i \;=\; \mathrm{f}(\mathrm{DIC}_i,\mathrm{TA}_i) \\ \mathrm{DIN}_i \;=\; (\mathrm{NO}_3^-)_i\ [+\ (\mathrm{NO}_2^-)_i + (\mathrm{NH}_4^+)_i\ ]\ (\text{nitrate [+ nitrite + ammonia]}) \\ \mathrm{DIP}_i \;=\; (\mathrm{HPO}_4^{2-})_i\ \text{and/or}\ (\mathrm{PO}_4^{3-})_i\ (\text{phosphate}) \\ \mathrm{CHL}a_i \;=\; (\text{chlorophyll}\ a)_i \\ \mathrm{PON}_i \;=\; (\mathrm{PhyN} + \mathrm{ZooN} + \mathrm{DetN})_i \\ \mathrm{POC}_i \;=\; (\mathrm{PhyC} + \mathrm{ZooC} + \mathrm{DetC} + \text{Ge-C})_i \end{array} \right)$$

1) Dynamical, process-based modelling to derive mass flux estimates



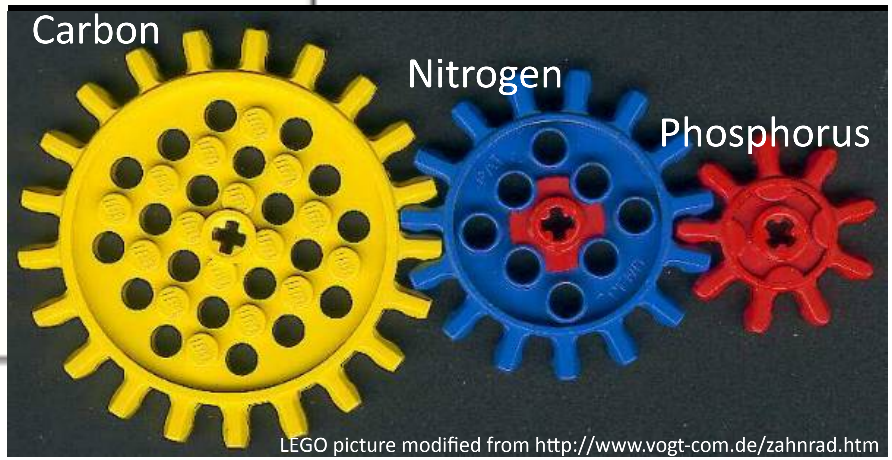
(Kreus et al. 2014, accepted, CSR)

1) Dynamical, process-based modelling to derive mass flux estimates



(Kreus et al. 2014)

After fitting model results to data we eventually derive annual budgets of **Carbon**-, **Nitrogen**-, & **Phosphorus** flux



2) Parameter estimation (some basics recalled)

→ the dynamical, process based model is used to derive mass flux estimates while explaining field observations

2) Parameter estimation (some basics recalled)

→ the dynamical, process based model is used to derive mass flux estimates while explaining field observations

The ever question is:
What is the appropriate model complexity (number of parameters & degree of freedom) for the data available?

or
How well are the model parameter values constrained by data?

2) Parameter estimation (some basics recalled)

The assimilation of data to estimate model parameters in order to come with some reliable model solution is an important aspect in marine biogeochemical-, and ecosystem modeling.

2) Parameter estimation (some basics recalled)

The assimilation of data to estimate model parameters in order to come with some reliable model solution is an important aspect in marine biogeochemical-, and ecosystem modeling.

Typically, the negative natural logarithm of the likelihood for minimization:

$$-log_e(L) = \mathrm{constant}_\Sigma + J^* = \mathrm{constant}_\Sigma + \sum_{i=1}^{N_i} \frac{1}{2} (\mathrm{y}_i^o - \mathrm{x}_i^o)^T R_i^{-1} (\mathrm{y}_i^o - \mathrm{x}_i^o)$$

$$= \mathrm{constant}_\Sigma + \frac{1}{2}\chi^2$$

or

$$\chi^2 = \mathrm{constant}_\Sigma - 2 \cdot log_e(L) = \sum_{i=1}^{N_i} (\mathrm{y}_i^o - \mathrm{x}_i^o)^T R_i^{-1} (\mathrm{y}_i^o - \mathrm{x}_i^o)$$

R_i: Covariance matrix of observations
N_i: Number of dates with available observations

3) What is meant by *parameter identifiability*?

We derived $J = \chi^2$ because it is the typical cost function used to discuss issues of parameter identifiability.

3) What is meant by *parameter identifiability*?

We derived $J = \chi^2$ because it is the typical cost function used to discuss issues of parameter identifiability.

General information: Parameter identifiability

a) is associated with the uncertainty of a parameter estimate
b) is concerned with the specification of a confidence region in parameter space
c) is a matter of model structure, experimental design, and data availability

3) What is meant by *parameter identifiability*?

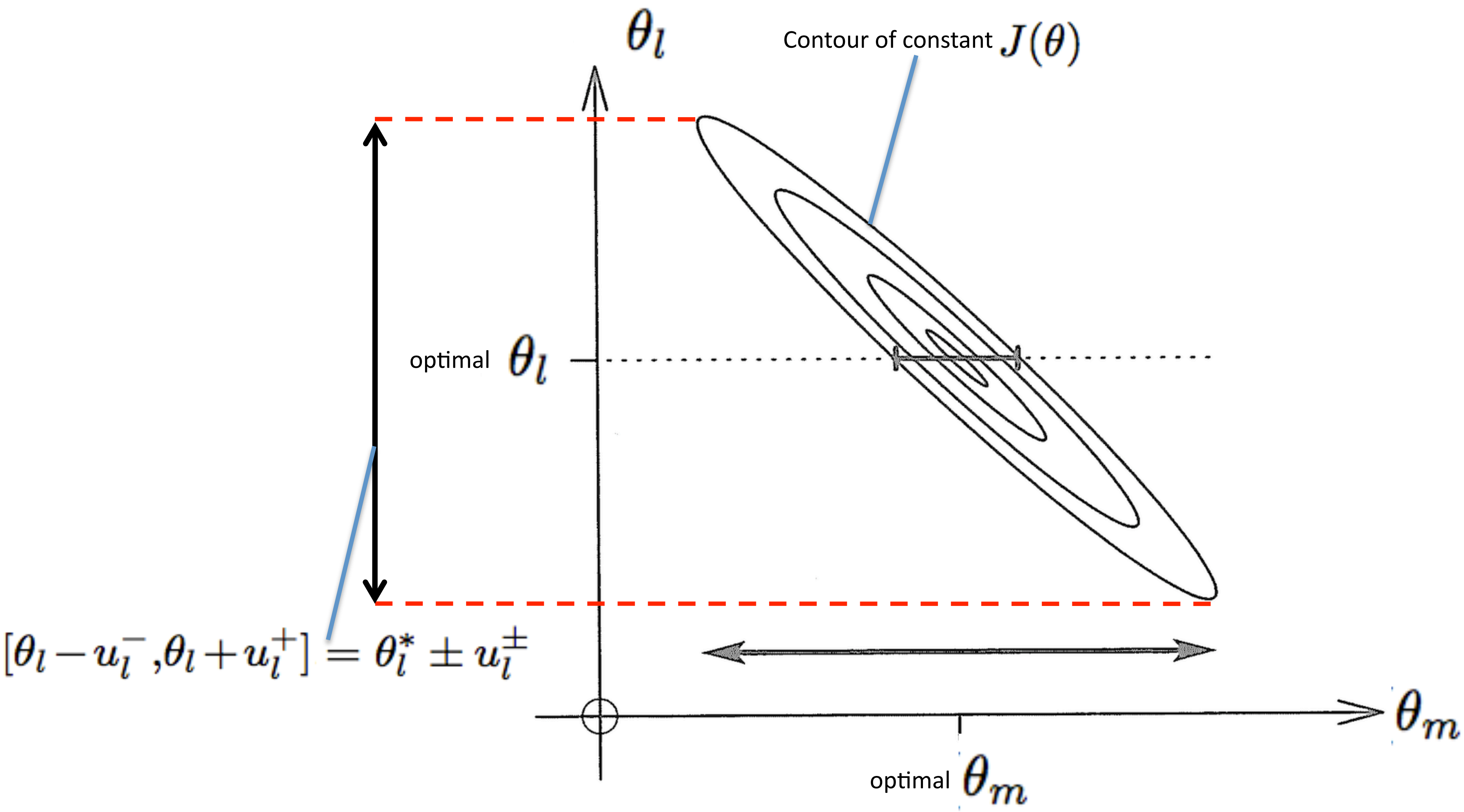
We derived $J = \chi^2$ because it is the typical cost function used to discuss issues of parameter identifiability.

General information: Parameter identifiability

a) is associated with the uncertainty of a parameter estimate
b) is concerned with the specification of a confidence region in parameter space
c) is a matter of model structure, experimental design, and data availability

→ to investigate whether a parameter is structurally and practically identifiable can also be seen as an intermediate step before deriving a posterior distribution where a prior is eventually included (e.g. for MCMC methods).

3) What is meant by *parameter identifiability*?



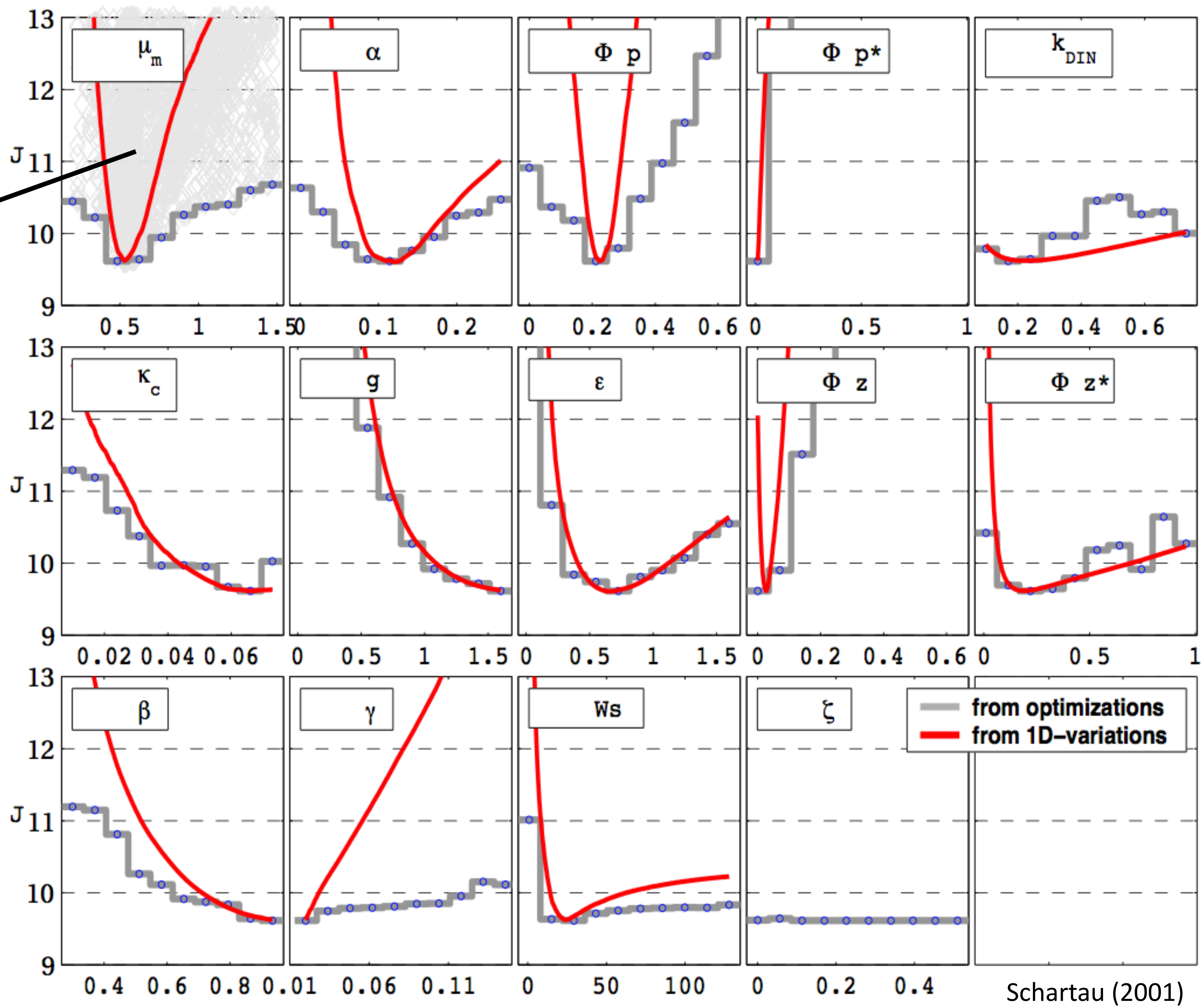
→ The contours (shape) of the cost function determine the error margins $[\theta_l - u_l^-, \theta_l + u_l^+]$ of the parameter estimate.

Figure modified from Sivia (with Skilling) Data Analysis: A Baysian Tutorial

3) What is meant by *parameter identifiability*?

Optimization with a micro-genetic algorithm

gray = values of J in the vicinity of the minimum during the search process



Schartau (2001)

3) What is meant by *parameter identifiability*?

→ The contours (shape) of the cost function determine the error margins $[\theta_l - u_l^-, \theta_l + u_l^+]$ of the parameter estimate.

→ A parameter is identifiable if the error margins are finite, or better, if the error margins are within a range of credible parameter values.

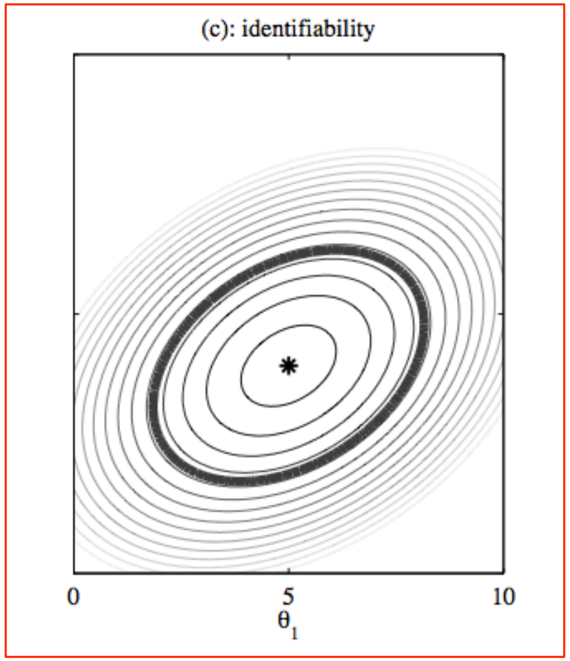
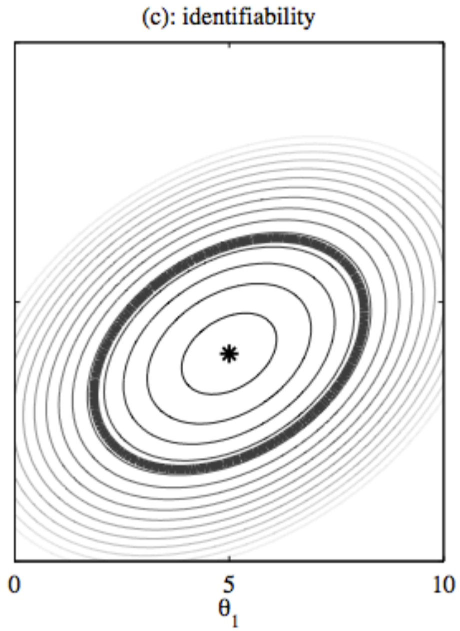


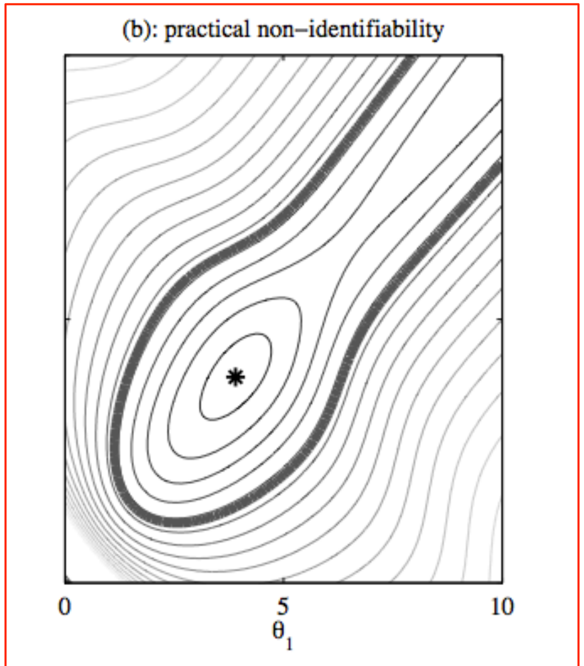
Figure modified from Raue et al. (2010, IET Systems Biology)

3) What is meant by *parameter identifiability*?

→ A parameter is identifiable if the error margins are finite, or better, if the error margins are within a range of credible parameter values.

→ A parameter is termed non-identifiable as soon as one of the error margins is infinite, or if the error margins are well outside a range of credible parameter values.

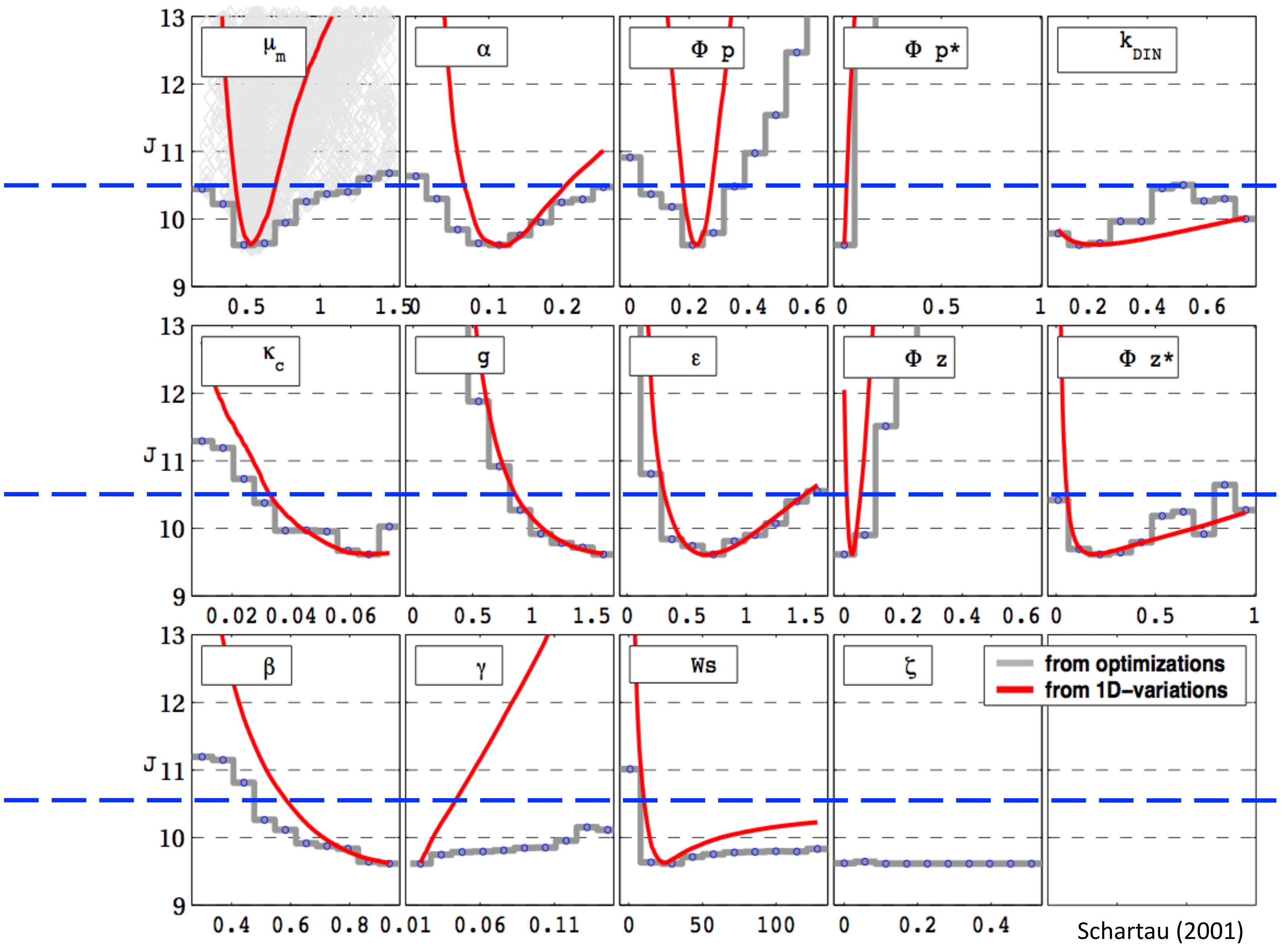




Figures modified from Raue et al. (2010, IET Systems Biology)

3) What is meant by *parameter identifiability*?

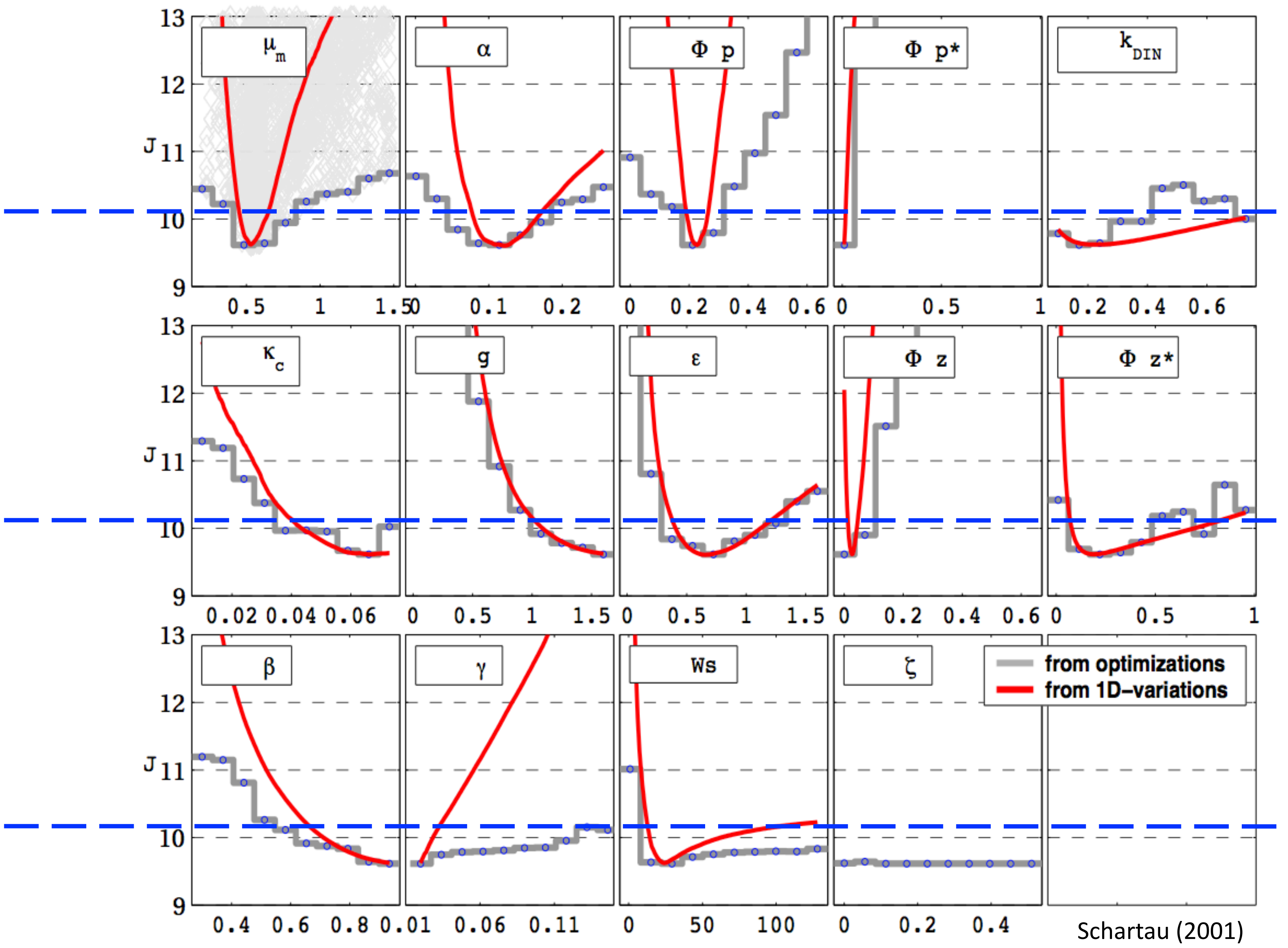
?



Schartau (2001)

3) What is meant by *parameter identifiability*?

?



Schartau (2001)

3) What is meant by *parameter identifiability*?

The question now is:
Which contour (cost function value) specifies these error margins?

3) What is meant by *parameter identifiability*?

The question now is:
Which contour (cost function value) specifies these error margins?

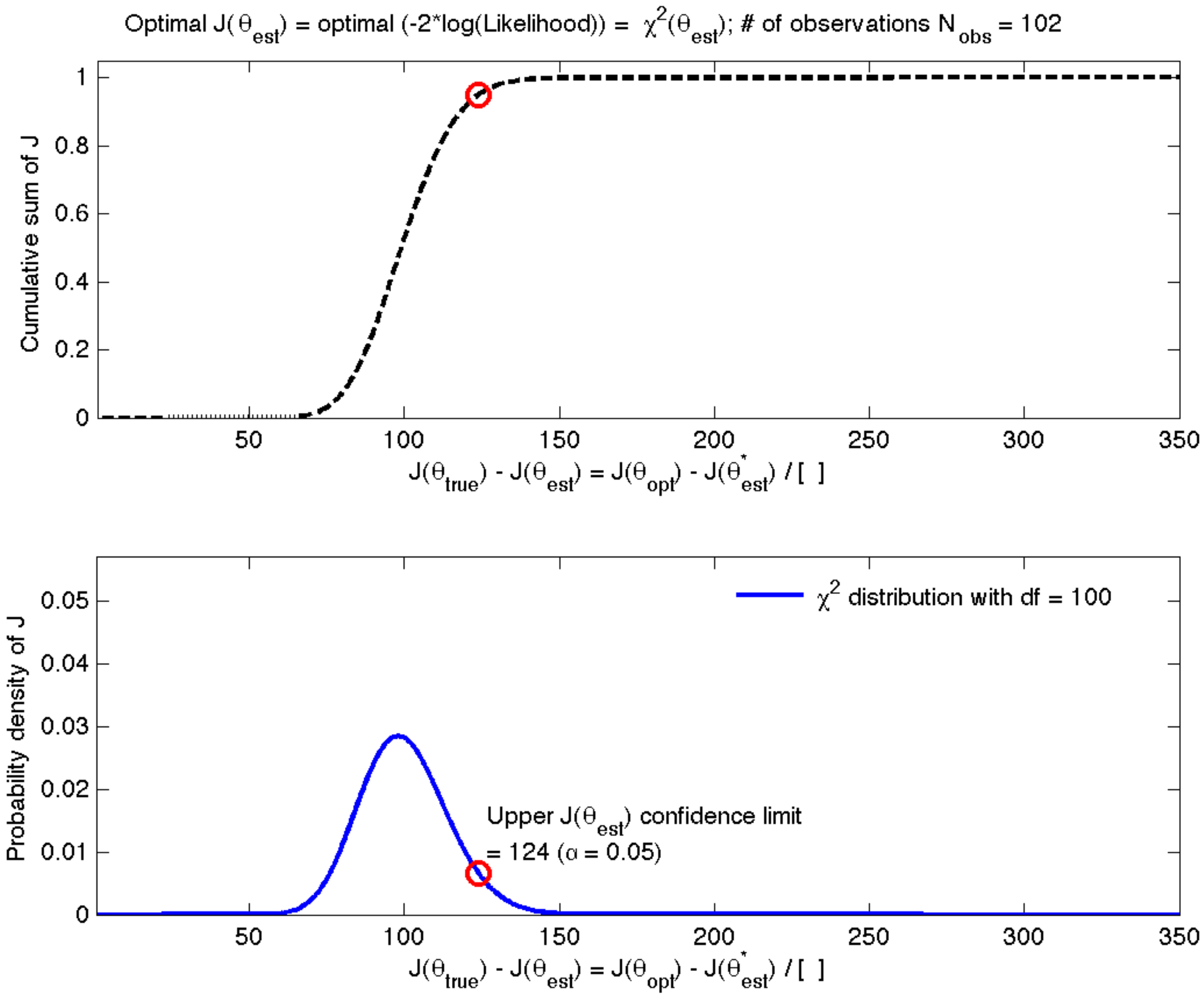
To answer this we need to find the contour of a confidence region that encloses the distribution (distance) of an optimal parameter estimate θ_l^* relative to the apparent true value θ_l^t:

$$(\theta_l^t - \theta_l^*) \approx \theta_l^* \pm u_l^\pm$$

According to the idealized theory we can specify e.g. the 95% confidence limit from a standard χ^2 -distribution for a given degree of freedom (df).
The degree of freedom is then equal to the number of observations minus the number of parameters of interest ($df = N_i - N_p$).

3) What is meant by *parameter identifiability*?

95% confidence limit of a standard χ^2 (df)-distribution with df = 100



3) What is meant by *parameter identifiability*?

In practice, with our non-linear model, the actual χ^2 (df) - distribution is expected to look different & we actually do not know the exact degree of freedom.

Question:
How can we derive a distribution that is representative for our cost function and helps us specifying our own confidence limits explicitly?

3) What is meant by *parameter identifiability*?

How can we derive a distribution that is representative for our cost function and specify our own confidence limits explicitly?

For an answer let us consider the following situation:

1) the experiment (or field observation) was repeated many times with similar conditions, comparable variability, & same type of data were collected (ideally at same dates)

2) the same model is used for parameter estimation

3) for every repeated experiment we retrieve new optimal parameter values

3) What is meant by *parameter identifiability*?

How can we derive a distribution that is representative for our cost function and specify our own confidence limits explicitly?

For an answer let us consider the following situation:

1) the experiment (or field observation) was repeated many times with similar conditions, comparable variability, & same type of data were collected (ideally at same dates)

2) the same model is used for parameter estimation

3) for every repeated experiment we retrieve new optimal parameter values

Then we can postulate:

$$(\theta_l^t - \theta_l^*) \approx (\theta_l^* - \theta_l^{r*}) = \theta_l^* \pm u_l^{\pm}$$

with θ_l^{r*} being an optimal parameter value retrieved from a repeated experiment

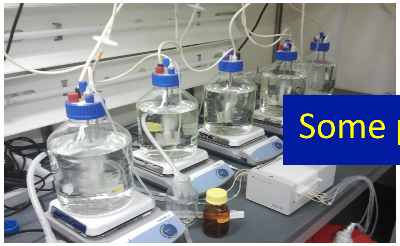
4) Resampling strategy to specify confidence regions in parameter space

In analogy to the standard χ^2 (*df*)-distribution we can define the confidence interval as a threshold Δ_α in the likelihood (Raue et al., 2010):

$$J(\theta^t) - J(\theta^*) \approx J(\theta^*) - J(\theta^{r*}) < \Delta_\alpha \leq \widehat{Q}$$

with our $\widehat{Q}$ as the 1-α quantile of our distribution of repeated optimizations (with repeated experimental data).

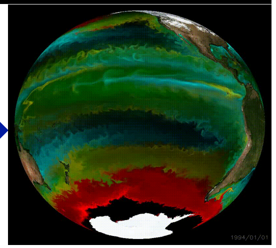
4) Resampling strategy to specify confidence regions in parameter space



Some prior information

Posterior error margins as prior

(M. Follows and colleagues, media library, MIT)



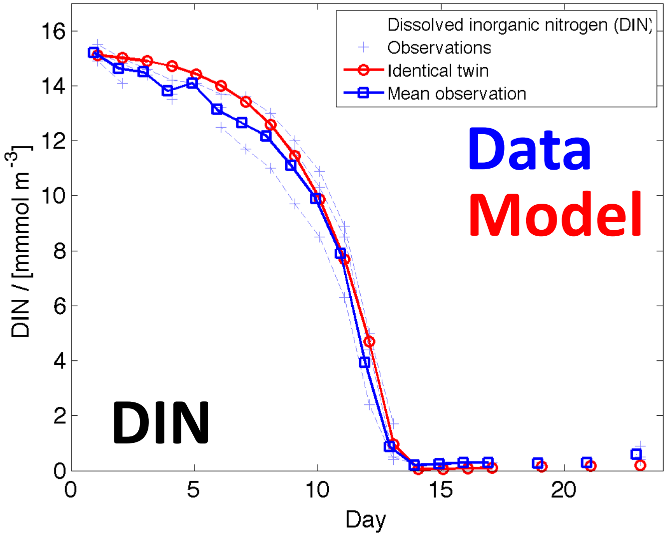
Laboratory conditions
→ range of credible parameter values

Field experiments
→ optimal parameter estimates + uncertainties

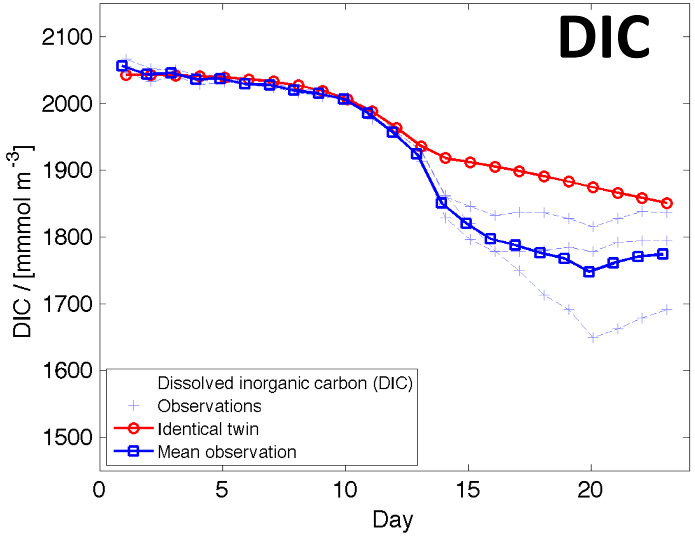
Example:
mesocosm experiments

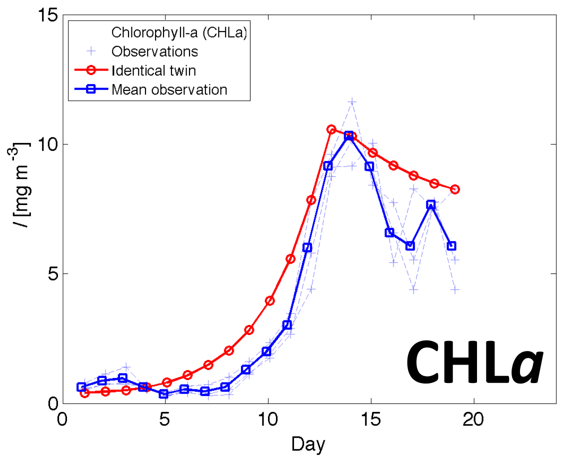
Large scale simulations
→ e.g. uncertainties in flux estimates based on uncertainties in parameter values

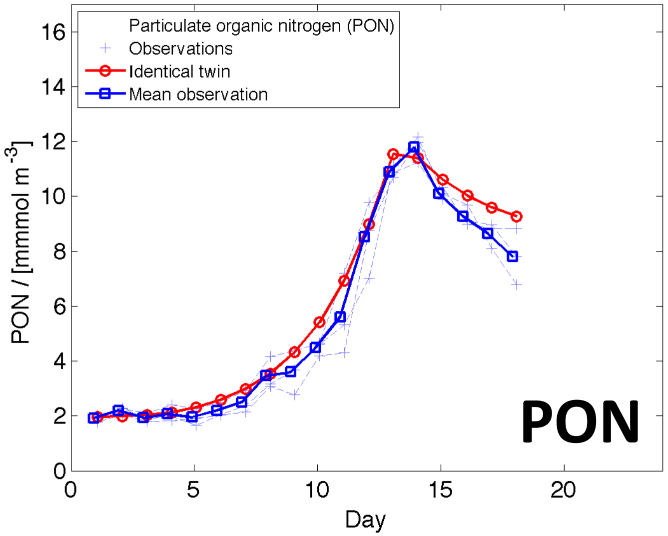
4) Resampling strategy to specify confidence regions in parameter space

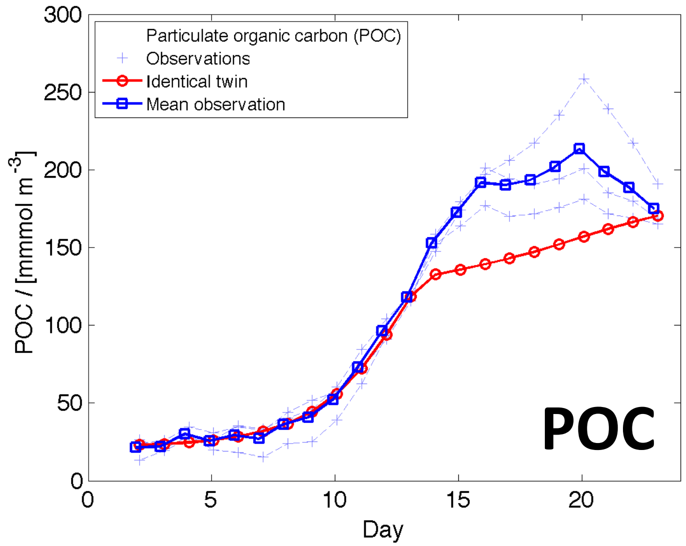


Three mesocosms with









4) Resampling strategy to specify confidence regions in parameter space

$$\chi^2 = \text{constant}_\Sigma - 2 \cdot log_e(L) = \sum_{i=1}^{N_i} (\mathrm{y}_i^o - \mathrm{x}_i^o)^T R_i^{-1} (\mathrm{y}_i^o - \mathrm{x}_i^o)$$

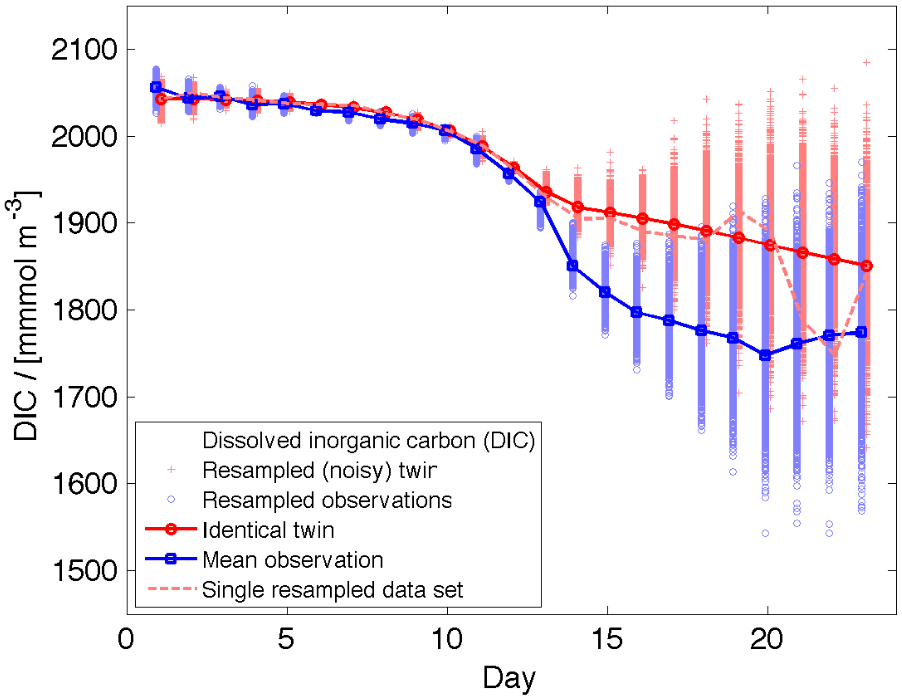
$$R_i = \begin{pmatrix} \sigma_{\mathrm{DIC}}^2 & \rho_{1,2}\sigma_{\mathrm{DIC}}\sigma_{\mathrm{DIN}} & \rho_{1,3}\sigma_{\mathrm{DIC}}\sigma_{\mathrm{CHLa}} & \rho_{1,4}\sigma_{\mathrm{DIC}}\sigma_{\mathrm{PON}} & \rho_{1,5}\sigma_{\mathrm{DIC}}\sigma_{\mathrm{POC}} \\ \rho_{1,2}\sigma_{\mathrm{DIC}}\sigma_{\mathrm{DIN}} & \sigma_{\mathrm{DIN}}^2 & . & . & . \\ \rho_{1,3}\sigma_{\mathrm{DIC}}\sigma_{\mathrm{CHLa}} & . & \sigma_{\mathrm{CHLa}}^2 & . & . \\ \rho_{1,4}\sigma_{\mathrm{DIC}}\sigma_{\mathrm{PON}} & . & . & \sigma_{\mathrm{PON}}^2 & . \\ \rho_{1,5}\sigma_{\mathrm{DIC}}\sigma_{\mathrm{POC}} & . & . & . & \sigma_{\mathrm{POC}}^2 \end{pmatrix}$$

→ Construct time-varying covariance matrix for resampling, according to error and correlation information of the observations

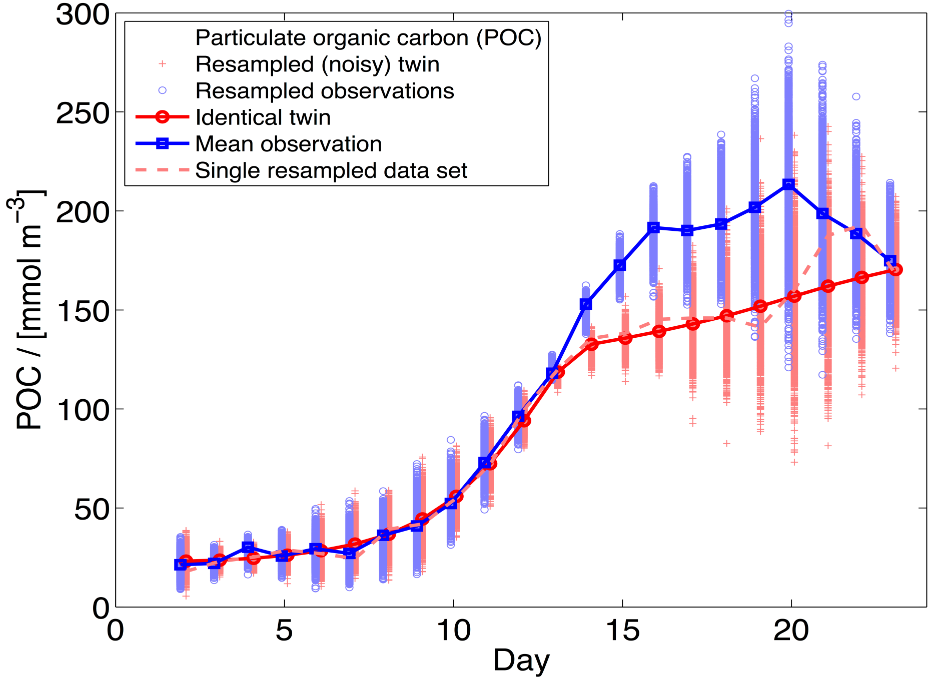
4) Resampling strategy to specify confidence regions in parameter space

Generate a series of resample set; here 2000

DIC (dissolved inorganic carbon)



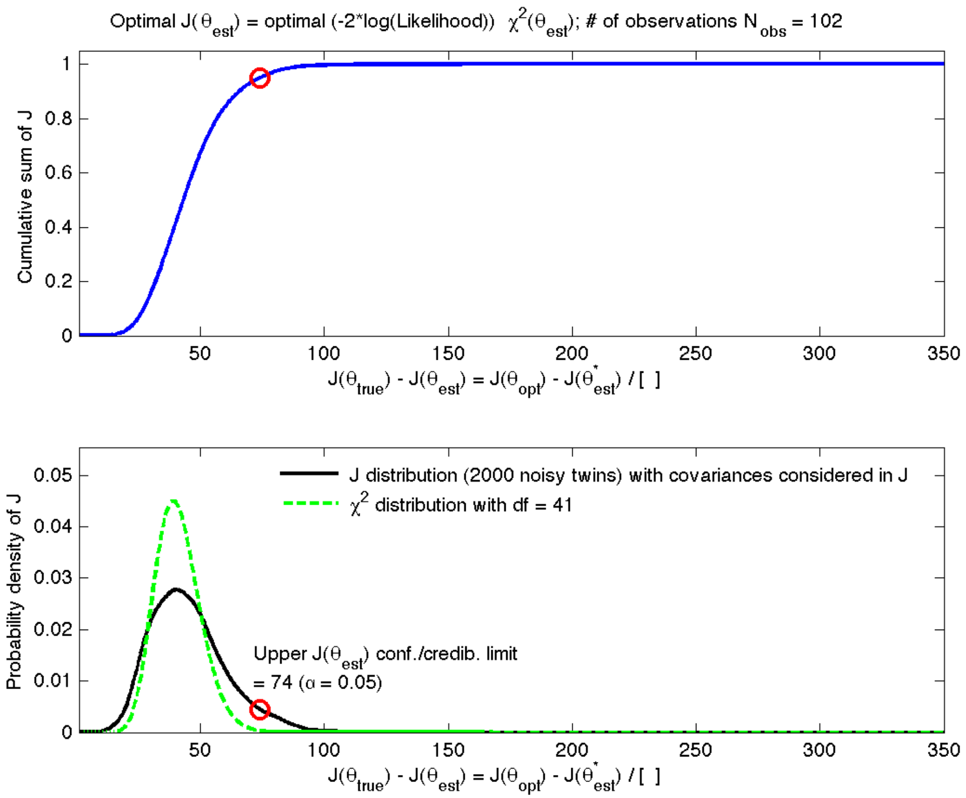
POC (particulate organic carbon)



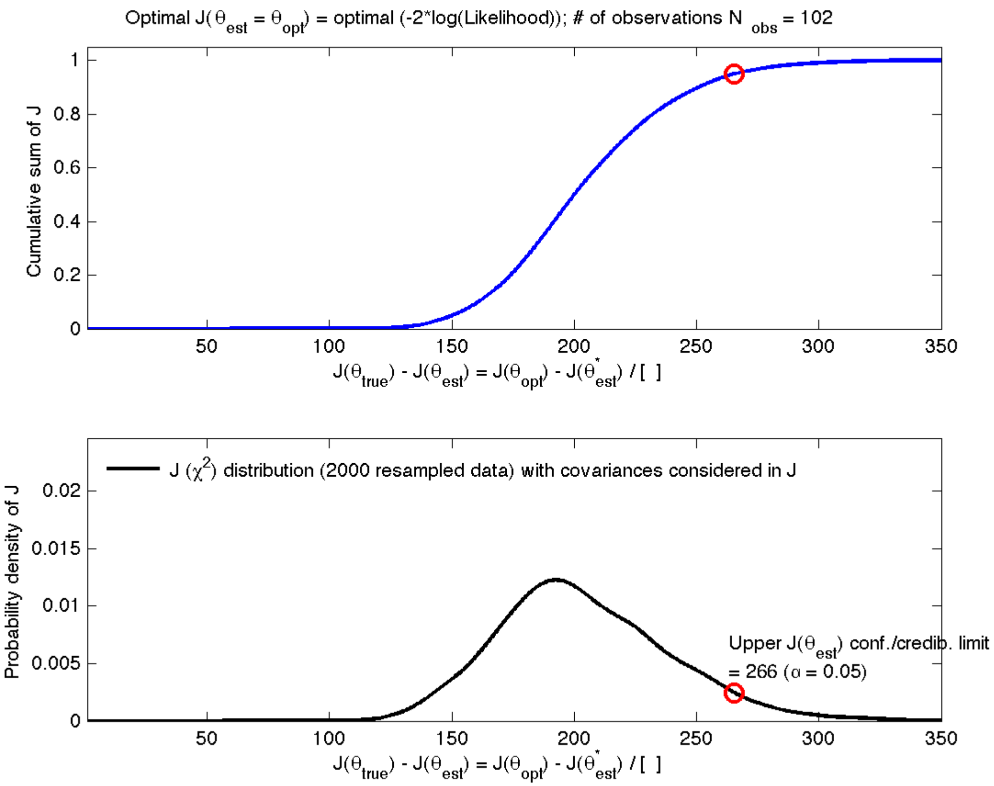
4) Resampling strategy to specify confidence regions in parameter space

Given the series of resample sets (here 2000) we then determine the distribution of cost function values.

TWIN (model resampled), J with covariances



DATA resampled, J with covariances

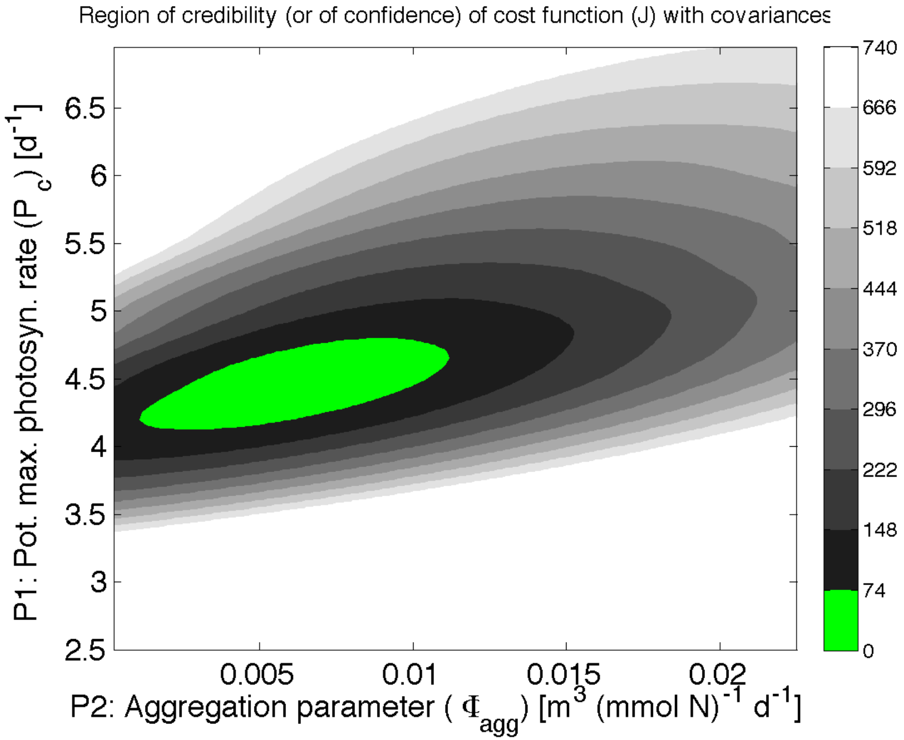


4) Resampling strategy to specify confidence regions in parameter space

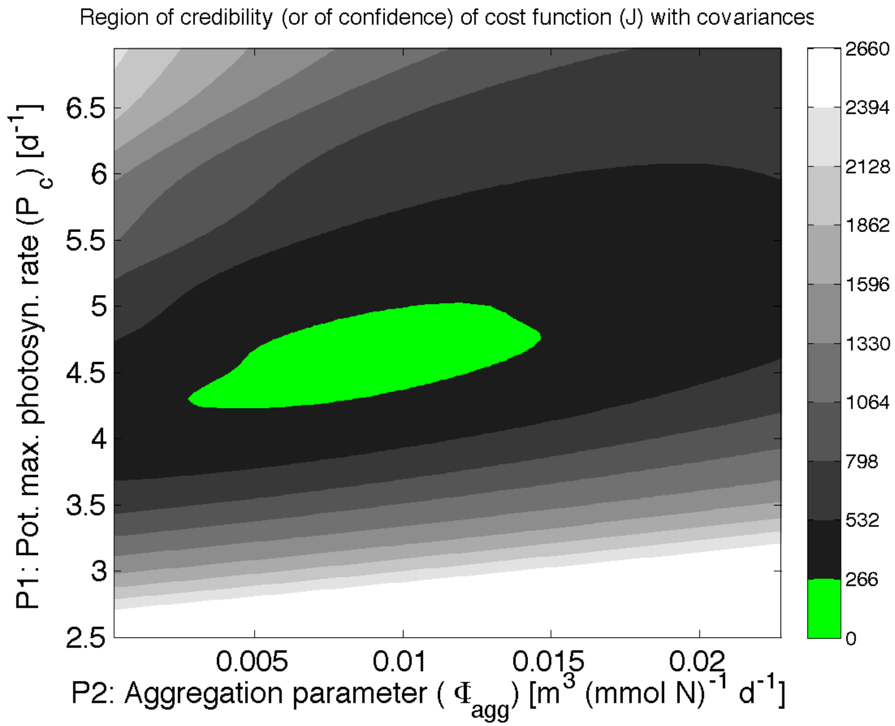
2D variations in parameter space

The good case

J with MODEL-TWIN and covariances



J with OBS-DATA and covariances



4) Resampling strategy to specify confidence regions in parameter space

Quadratic approximation:
Deriving error margins from the inversion of the Hessian matrix

Taylor series of the cost function at optimal parameter θ^*

$$J(\theta) = J(\theta^*) + (0) + \frac{1}{2}\sum_{l=1}^{N_p}\sum_{m=1}^{N_p} \underbrace{\left.\frac{\partial^2 J}{\partial\theta_l \partial\theta_m}\right|_{\theta=\theta^*}}_{\text{its inverse is related to } u_l^{\pm}} (\theta_l - \theta_l^*)(\theta_m - \theta_m^*) + \ldots$$

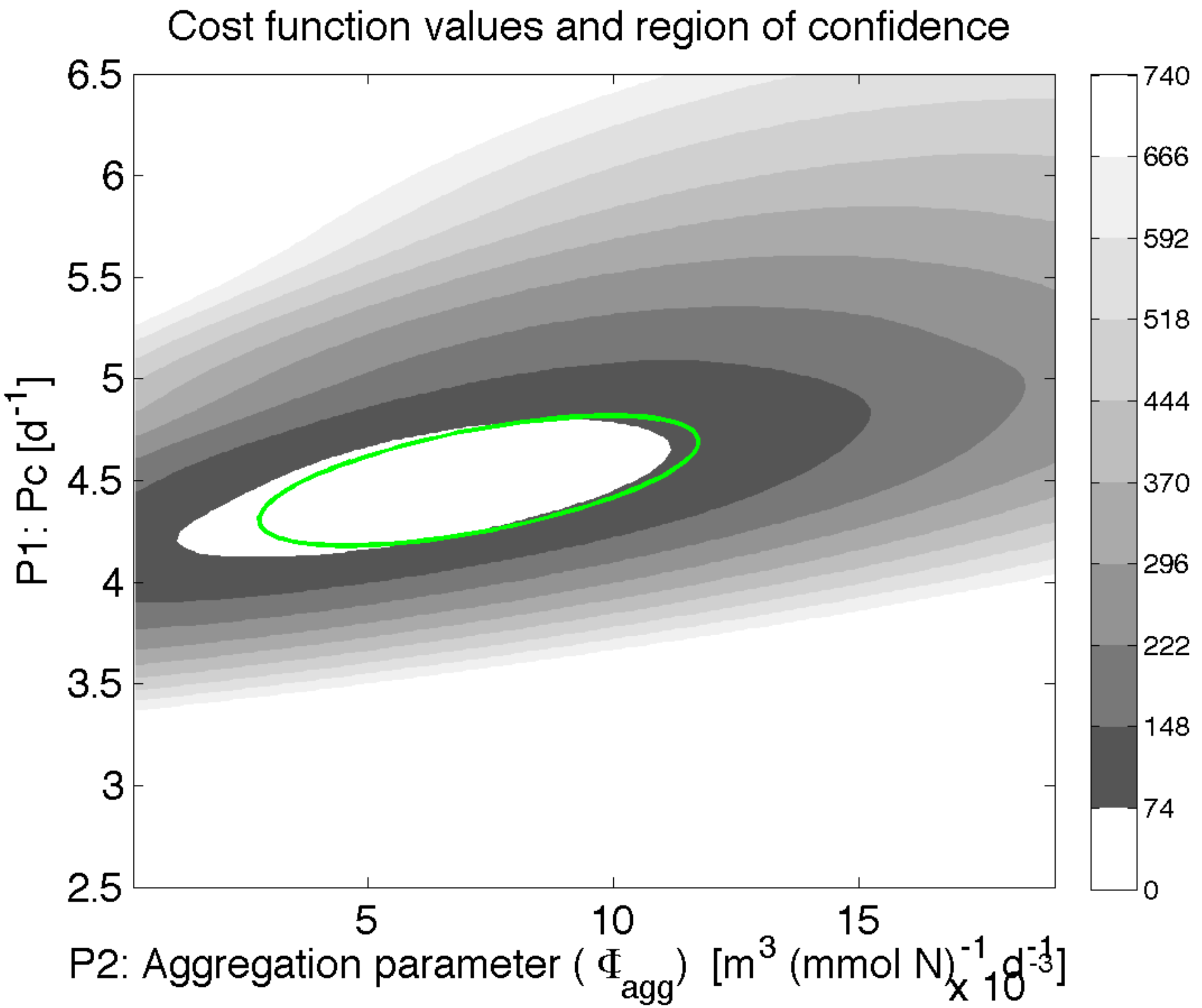
with $\mathcal{H}_{ll} = \frac{\partial^2 J}{\partial\theta_l^2}$ and $\mathcal{H}_{lm} = \frac{\partial^2 J}{\partial\theta_l \partial\theta_m}$ Hessian matrix $(N_p \times N_p)$

using its inverse $B = \mathcal{H}^{-1}$
we can derive the error margins for any Q specify from our J_{df}- distribution

$$u_l^{\pm} = \sqrt{Q \cdot B_{ll}}$$

4) Resampling strategy to specify confidence regions in parameter space

2D error ellipse derived from inverted approximated Hessian matrix

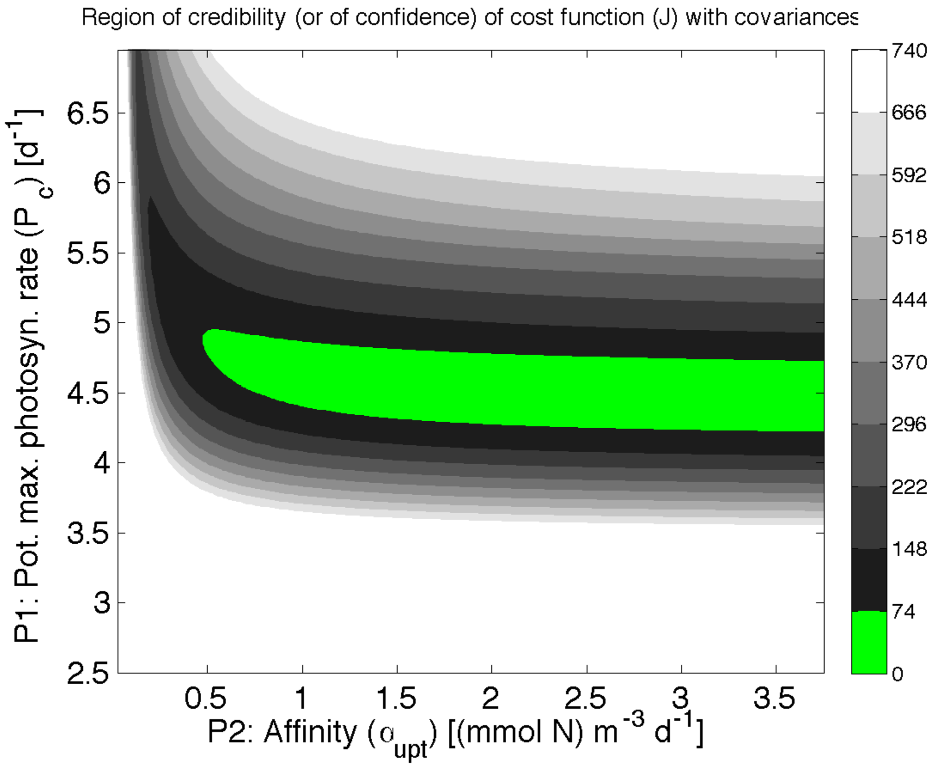


4) Resampling strategy to specify confidence regions in parameter space

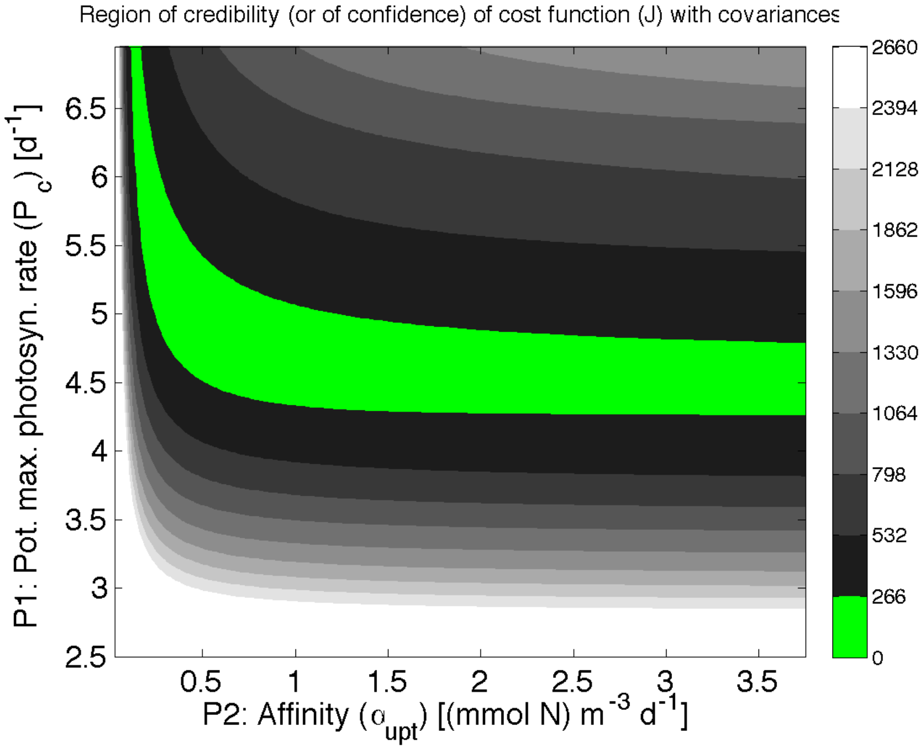
2D variations in parameter space

The not so good case

J with MODEL-TWIN and covariances

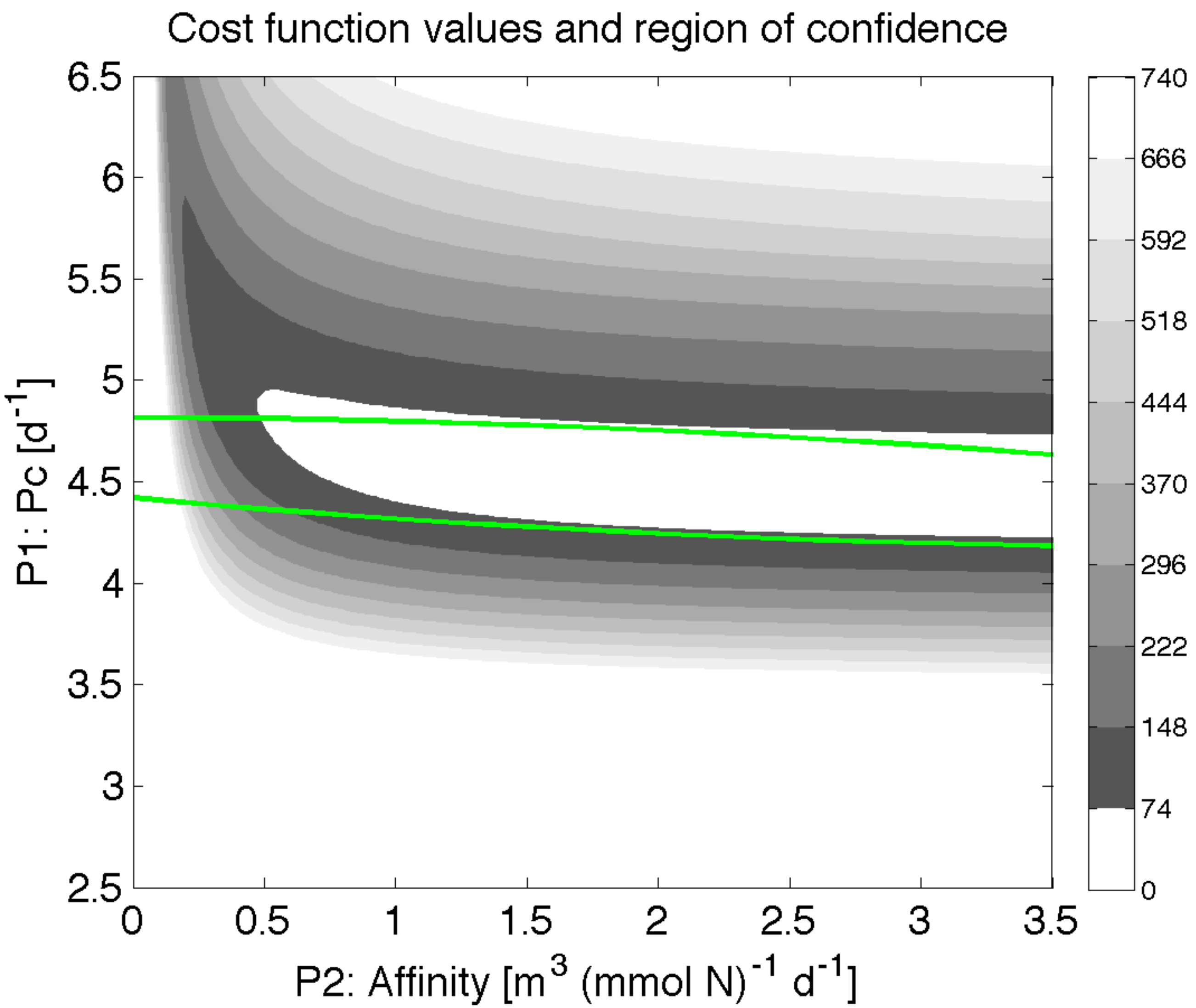


J with OBS-DATA and covariances



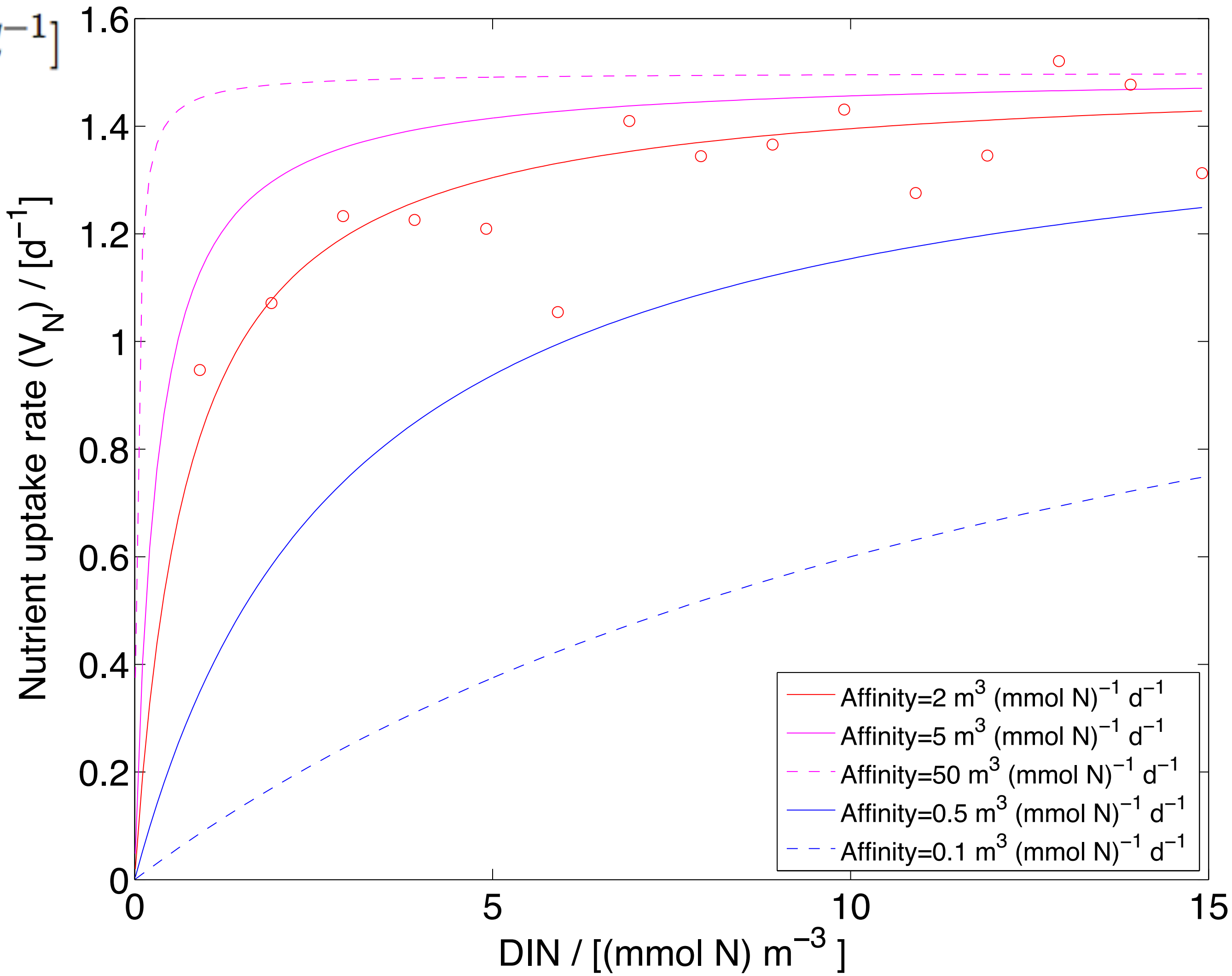
4) Resampling strategy to specify confidence regions in parameter space

2D error ellipse derived from inverted approximated Hessian matrix



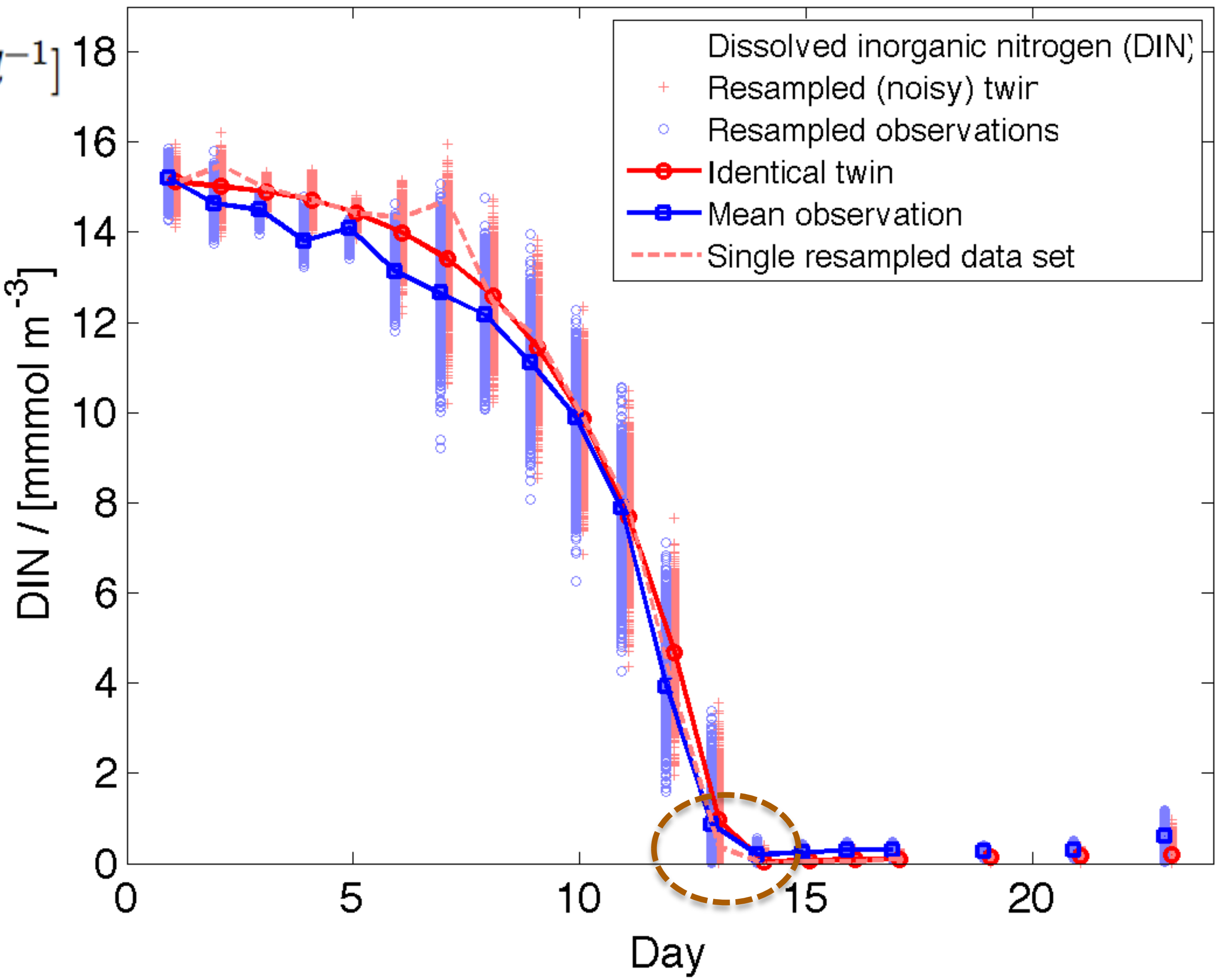
4) Resampling strategy to specify confidence regions in parameter space

$$V = \frac{V_{\max} \cdot DIN}{\left(\frac{V_{\max}}{\mathrm{A_{ffinity}}} + DIN\right)}; \quad [d^{-1}]$$



4) Resampling strategy to specify confidence regions in parameter space

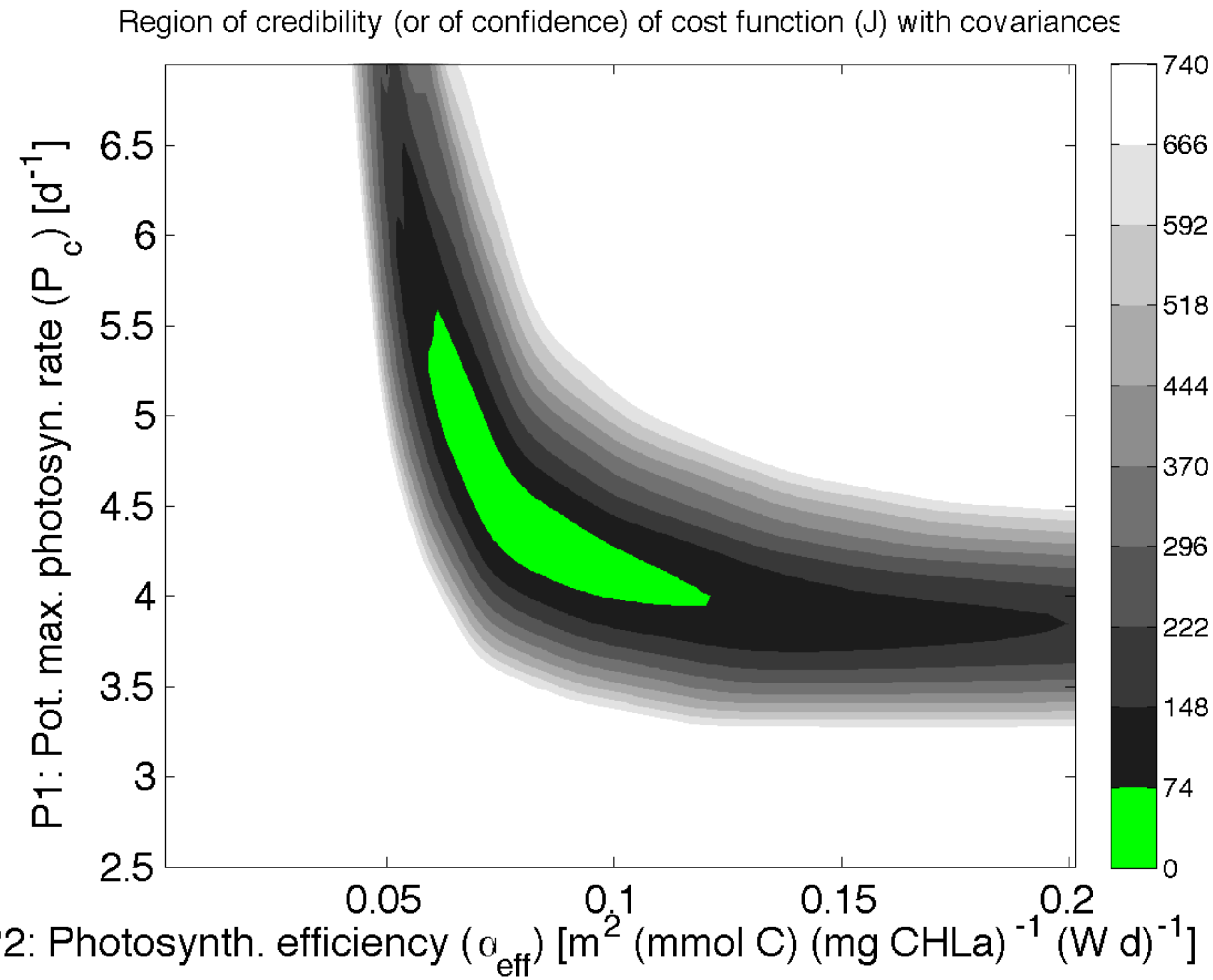
$$V = \frac{V_{\max} \cdot DIN}{\left(\frac{V_{\max}}{\mathrm{A_{ffinity}}} + DIN\right)}; \quad [d^{-1}]$$



4) Resampling strategy to specify confidence regions in parameter space

2D variations in parameter space

The banana case
(to be looked at
during the exercises)



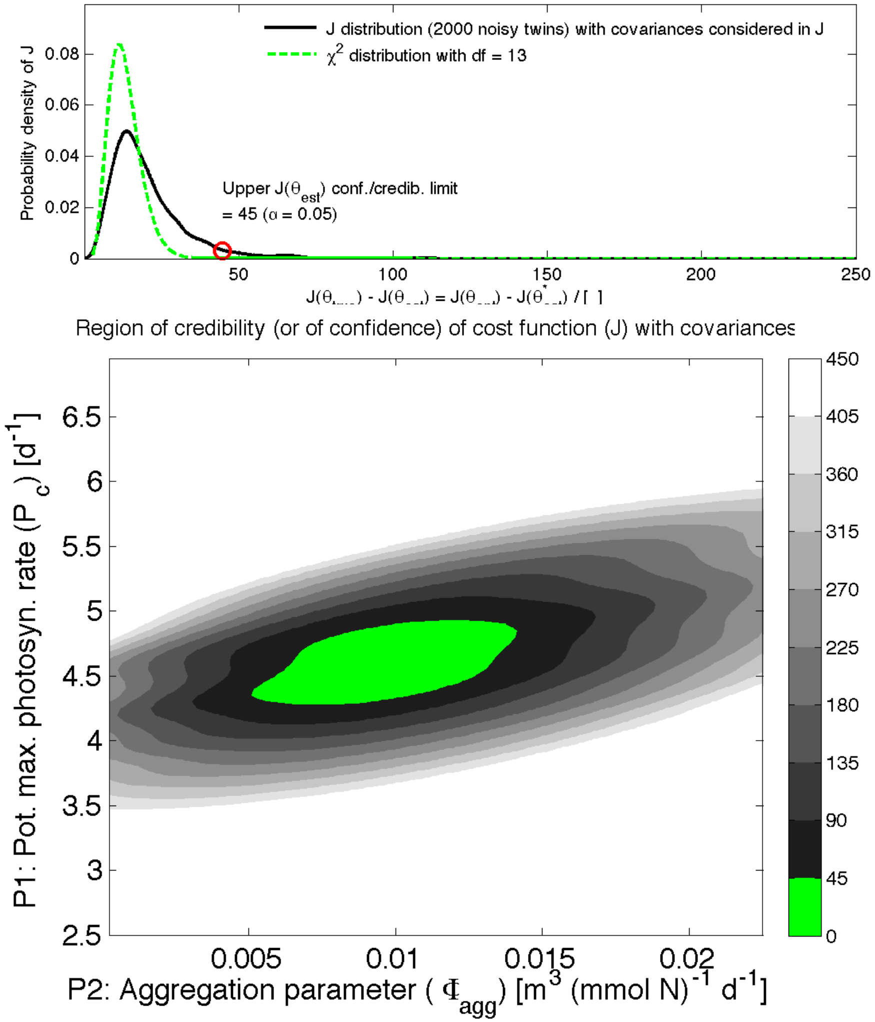
4) Resampling strategy to specify confidence regions in parameter space

Finally, let us assume that we had (e.g. remote sensing) chlorophyll *a* data only

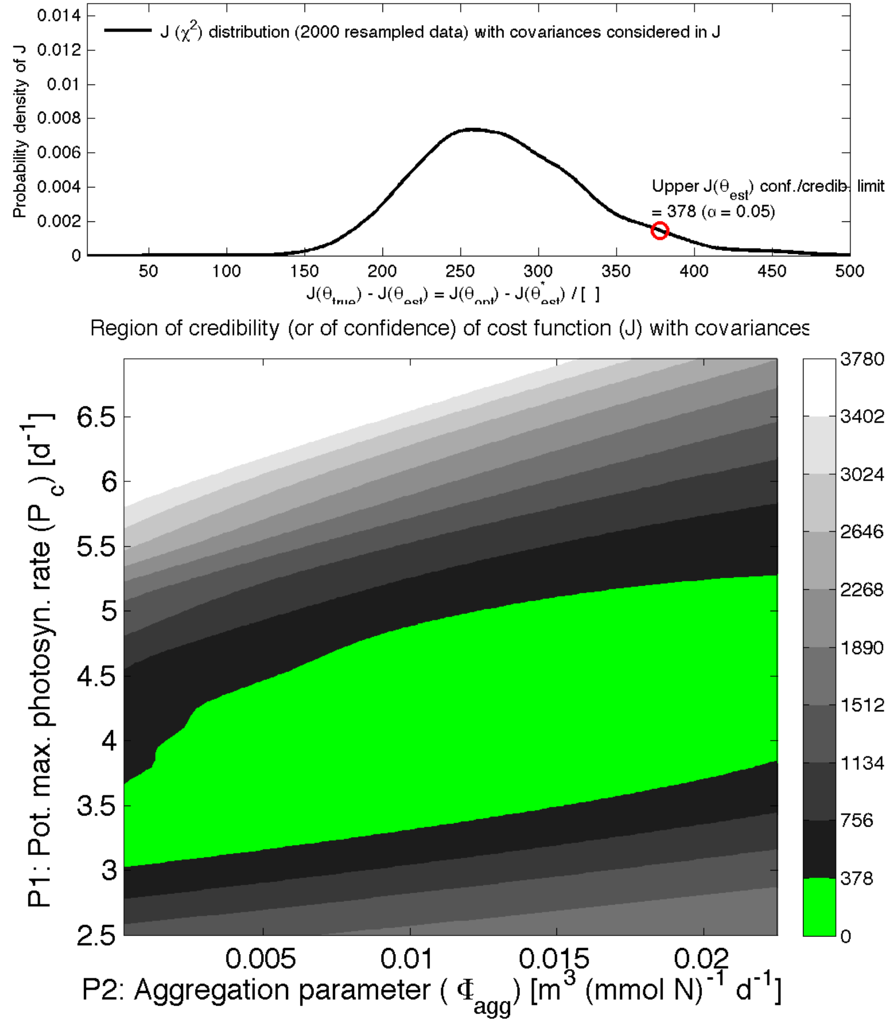
4) Resampling strategy to specify confidence regions in parameter space

Finally, let us assume that we had (e.g. remote sensing) chlorophyll *a* data only

J with MODEL-TWIN with CHL*a* only



J with OBS-DATA with CHL*a* only



Parameter identifiability of marine ecosystem models

Summary

1) Data is needed to constrain model parameters and to determine a unique model solution that provides reliable mass flux estimates; fits of model trajectories to data alone do not automatically yield optimal mass flux.

2) To determine error margins of an optimal parameter estimate and to assess whether a model parameter is well constrained depends on the confidence limits imposed.

3) The effective degree of freedom (df) is often unknown and therefore a standard χ^2 - distribution with prescribed df may not be helpful to set confidence limits.

4) Resampling (mimicing repeated experiments) can help to determine confidence limits explicitly and thus specify the error margins in parameter space

5) Error margins can then be derived, e.g. when combining the inverse Hessian with these confidence limits

6) The quadratic approximation is useful but has its limitations, in particular for parameters that describe saturation functions (e.g. similar to Michaelis-Menten)

2) Parameter estimation (some basics recalled)

Bayes' theorem (model parameter estimates, θ):

$$\mathrm{prob}(\theta^*|\mathrm{y}^o,\mathrm{x}) = \mathrm{prob}(\mathrm{y}^o|\theta^*,\mathrm{x}) \cdot \frac{\mathrm{prob}(\theta^*|\mathrm{x})}{\mathrm{prob}(\mathrm{y}^o|\mathrm{x})}$$

Posterior probability of best model parameter estimate (θ^*)

= likelihood · prior / evidence

2) Parameter estimation (some basics recalled)

Bayes' theorem (model parameter estimates, θ):

$$\mathrm{prob}(\theta^*|\mathrm{y}^o,\mathrm{x}) = \mathrm{prob}(\mathrm{y}^o|\theta^*,\mathrm{x}) \cdot \frac{\mathrm{prob}(\theta^*|\mathrm{x})}{\mathrm{prob}(\mathrm{y}^o|\mathrm{x})}$$

Posterior probability of best model parameter estimate (θ^*)

= likelihood · prior / evidence

If we consider one data set and one particular model, the denominator (evidence) remains insensitive to parameter variations[1]. We can therefore state:

$$\mathrm{prob}(\theta^*|\mathrm{y}^o,\mathrm{x}) \propto \mathrm{prob}(\mathrm{y}^o|\theta^*,\mathrm{x}) \cdot \mathrm{prob}(\theta^*|\mathrm{x})$$

Posterior probability of model parameter estimate is

proportional to likelihood · prior

[1] It does, however become relevant, if we are to compare between different models (e.g. of different complexity) to explain the same data. This particular aspect is not addressed here.

For multivariate Gaussian probability distributions we can write:

$$\mathrm{prob}(\theta^*|\mathrm{y}^o,\mathrm{x}) \propto \prod_{i=1}^{N_i} \frac{\exp\left[-\frac{1}{2}(\mathrm{y}_i^o - \mathrm{x}_i^o)^T R_i^{-1}(\mathrm{y}_i^o - \mathrm{x}_i^o)\right]}{\sqrt{(2\pi)^{N_j}\det(R_i)}}$$

$$\cdot \prod_{l=1}^{N_p} \frac{\exp\left[-\frac{1}{2}(\theta^g - \theta)^T B_\theta^{-1}(\theta^g - \theta)\right]}{\sqrt{(2\pi)^{N_p}\det(B_\theta)}}$$

R_i: Covariance matrix of observations
N_i: Number of dates with available observations

θ^g: First guess parameter values (prior information)
B_θ: Uncertainty (covariance) information of first guess parameter values
N_p: Number of parameters

2) Parameter estimation (some basics recalled)

Let us consider a flat prior (no information): $(\mathrm{prob}(\theta^*|\mathrm{x}) = 1)$

This way our posterior further reduces to:

$$\mathrm{prob}(\theta^*|\mathrm{y}^o, \mathrm{x}) \propto \prod_{i=1}^{N_i} \frac{\exp\left[-\frac{1}{2}(\mathrm{y}_i^o - \mathrm{x}_i^o)^T R_i^{-1} (\mathrm{y}_i^o - \mathrm{x}_i^o)\right]}{\sqrt{(2\pi)^{N_j} \det(R_i)}}$$

Here the denominator does not depend on the model parameters.
It is a constant:

$$\mathrm{prob}(\theta^*|\mathrm{y}^o, \mathrm{x}) \propto \mathrm{constant}_{\Pi} \cdot \prod_{i=1}^{N_i} \exp\left[-\frac{1}{2}(\mathrm{y}_i^o - \mathrm{x}_i^o)^T R_i^{-1} (\mathrm{y}_i^o - \mathrm{x}_i^o)\right] = L$$

Only the likelihood (L) is left!

In practice, rather than maximizing the likelihood distribution we consider the negative natural logarithm of the likelihood and minimize the non-constant term:

$$-log_e(L) = \text{constant}_\Sigma + J^* = \text{constant}_\Sigma + \sum_{i=1}^{N_i} \frac{1}{2} (\mathrm{y}_i^o - \mathrm{x}_i^o)^T R_i^{-1} (\mathrm{y}_i^o - \mathrm{x}_i^o)$$

$$= \text{constant}_\Sigma + \frac{1}{2}\chi^2$$

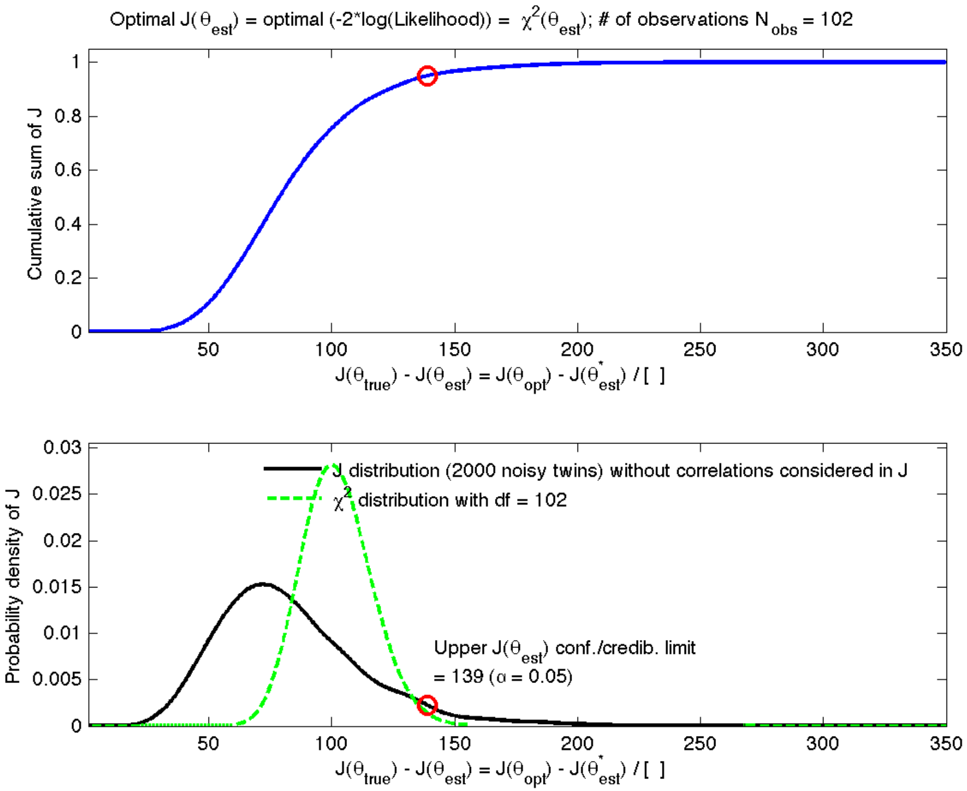
or

$$\chi^2 = \text{constant}_\Sigma - 2 \cdot log_e(L) = \sum_{i=1}^{N_i} (\mathrm{y}_i^o - \mathrm{x}_i^o)^T R_i^{-1} (\mathrm{y}_i^o - \mathrm{x}_i^o)$$

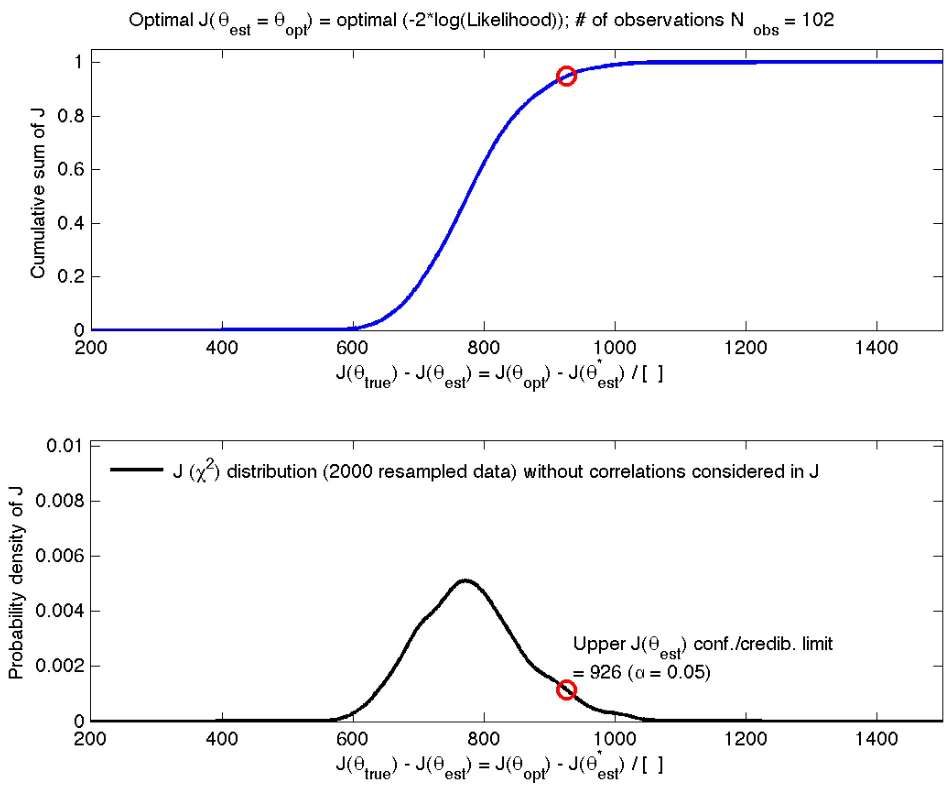
R_i: Covariance matrix of observations
N_i: Number of dates with available observations

Given the series of resample sets (here 2000) we then determine the distribution of cost function values.

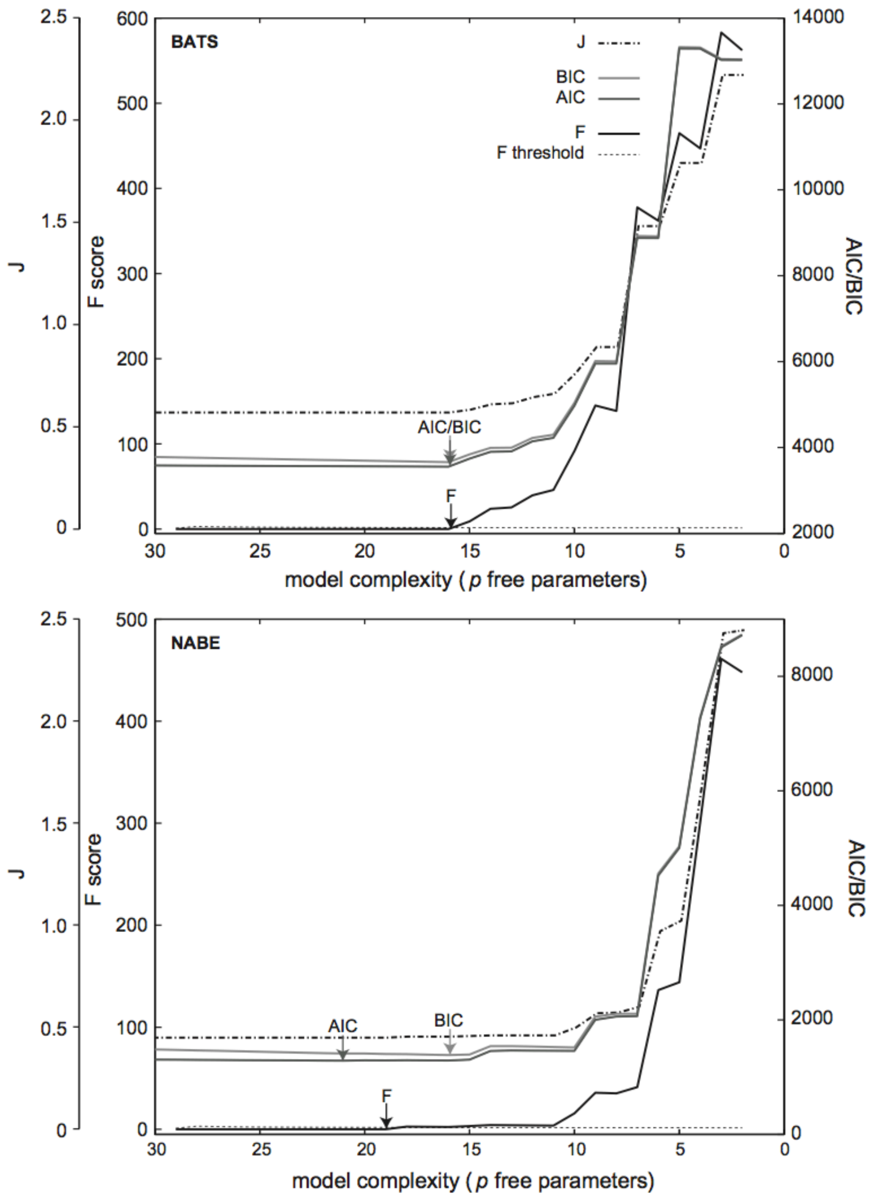
TWIN (model resampled), J without covariances

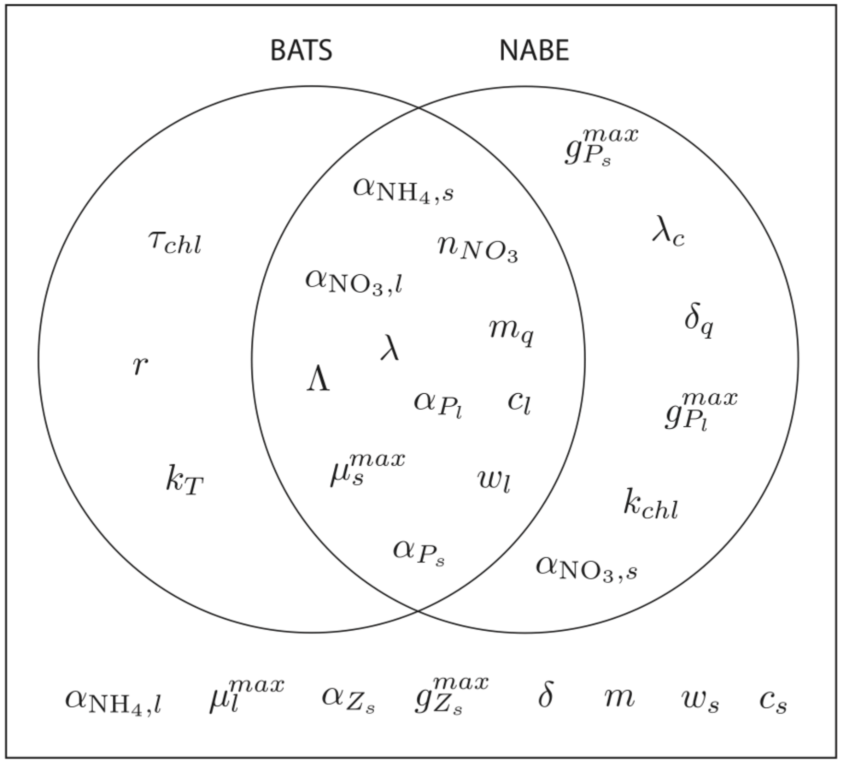


DATA resampled, J without covariances



Reducing parameter space without degrading model fit to data





Ward et al., (2013)

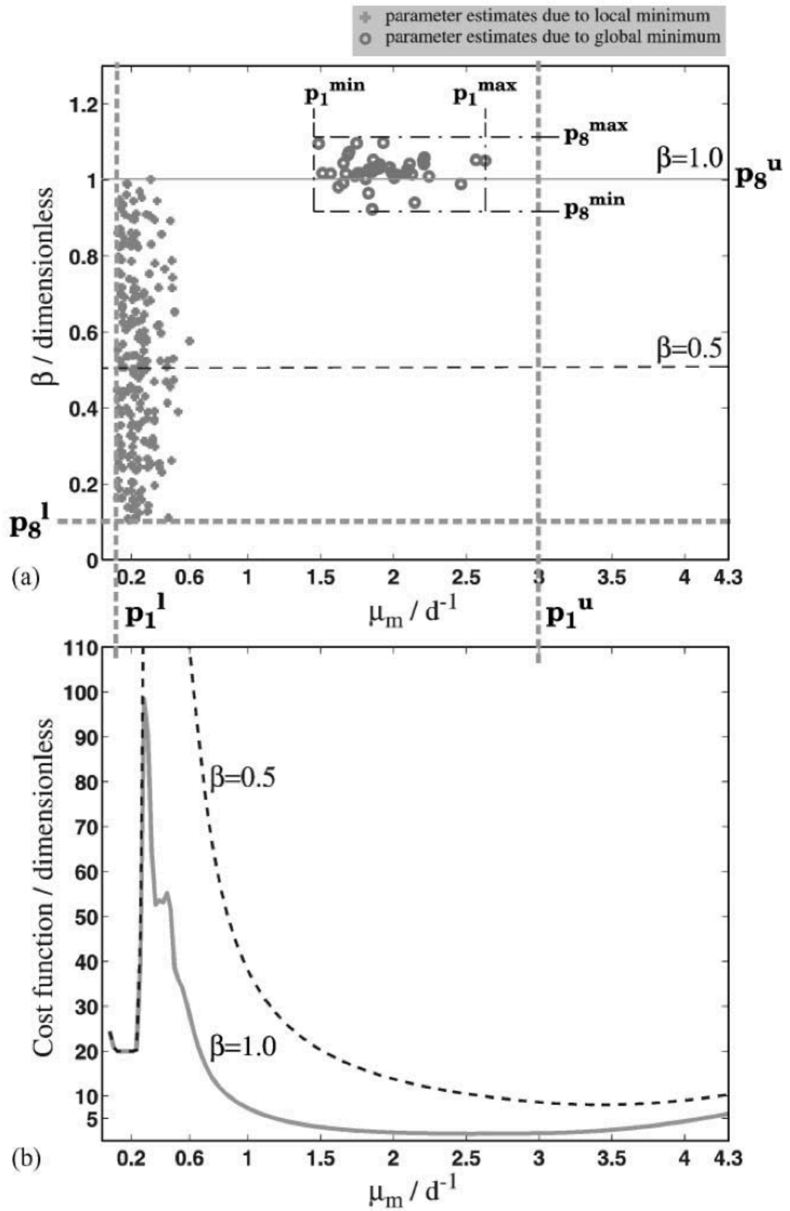
F-statistics for model reduction

$$F = \left[\frac{J_{red} - J_{full}}{J_{full}}\right] \left[\frac{p_{full} - p_{red}}{N - p_{full}}\right]^{-1}$$

$$F = \left[\frac{\text{increased error}}{\text{error in full model}}\right] \left[\frac{\text{increased parsimony}}{\text{parsimony of full model}}\right]^{-1}$$

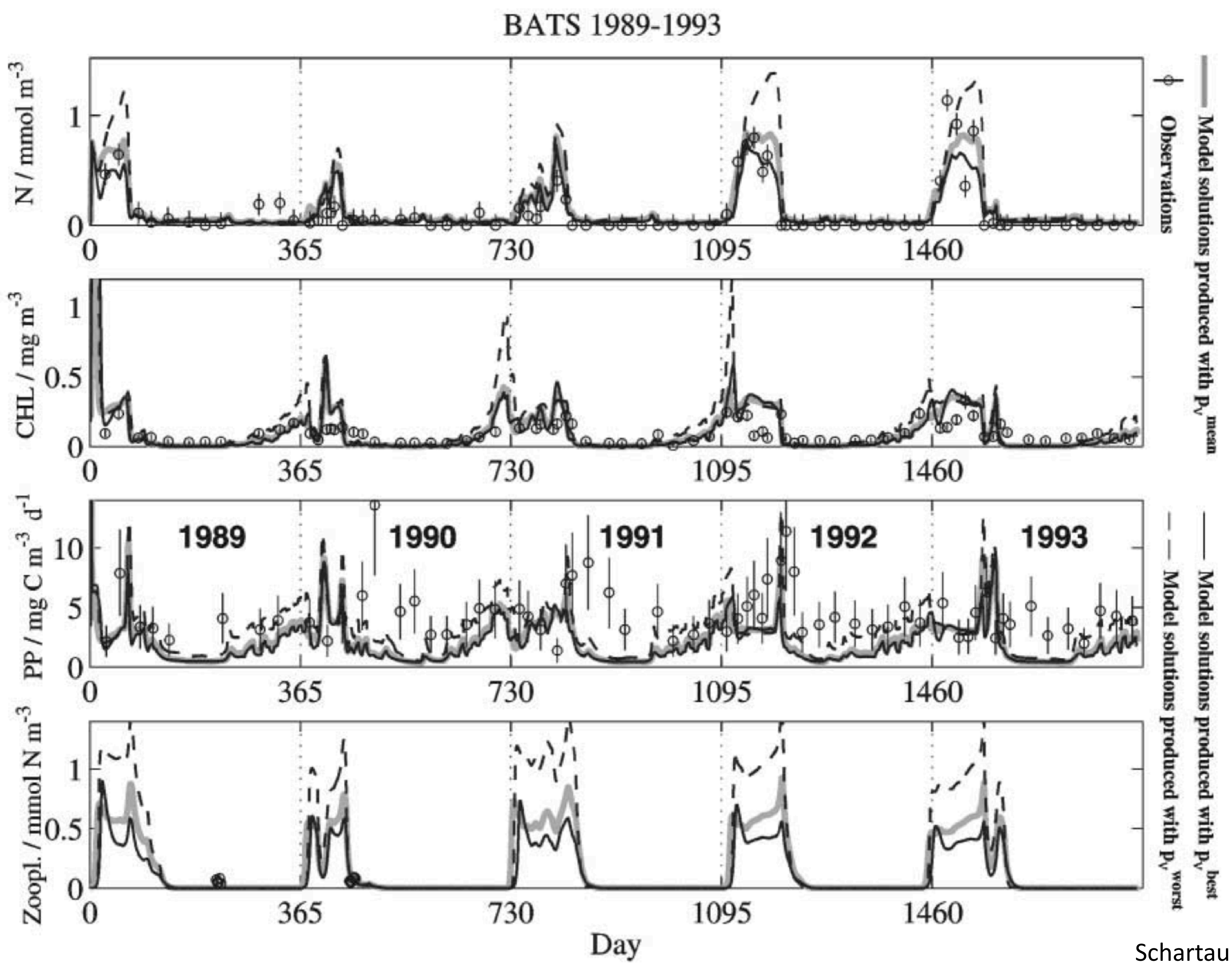
Ward et al., (2013)

Local minima with Two distinct model solutions at BATS site



Schartau et al., (2001)

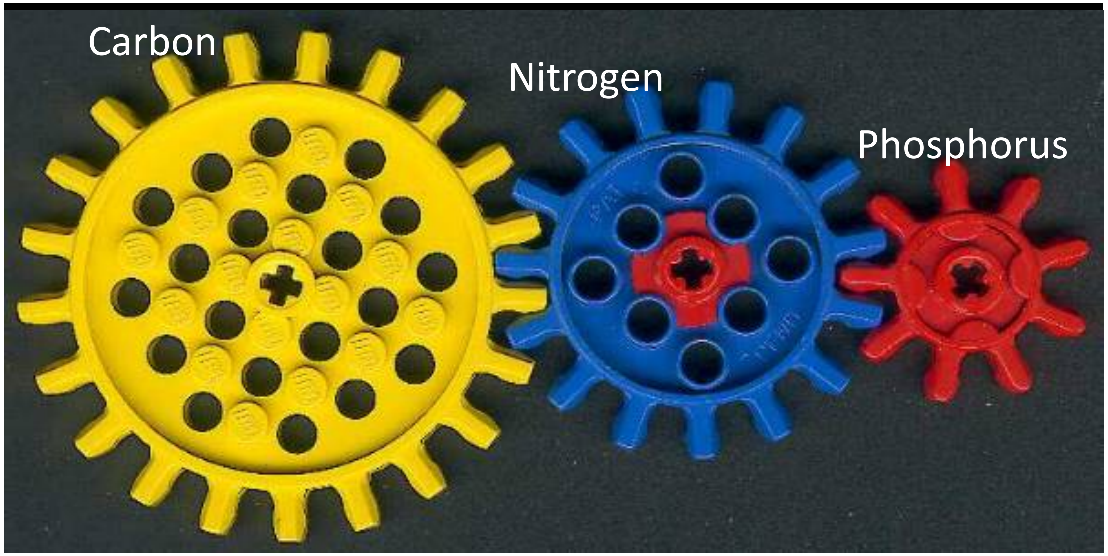
Two distinct model solutions at BATS site (nearly same cost function values)



Schartau et al., (2001)

Annual total production (**TP**), remineralisation (**RM**) & export ratio (**pe-ratio**) upper 60 meters

	TP	**RM**	**EX**	**pe-ratio**$=\frac{EX}{TP}$
	mol m^{-2} a^{-1}	mol m^{-2} a^{-1}	mol m^{-2} a^{-1}	n.d.
A: variable $q_{N:P}$				
Carbon (C):	14.16 ± 0.71	10.21 ± 0.55	3.8 ± 0.26	0.27 ± 0.01
Nitrogen (N):	1.40 ± 0.07	1.08 ± 0.06	0.31 ± 0.02	0.22 ± 0.01
Phosphorus (P):	0.114 ± 0.005	0.097 ± 0.005	0.016 ± 0.001	0.14 ± 0.01
C:N:P:	125:12:1	105:11:1	245:20:1	



LEGO picture modified from http://www.vogt-com.de/zahnrad.htm

Preferential remineralisation of phosphorus (P) within upper productive layer

Modified from Kreus et al., (2014, accepted)